

The American Economic Review

103

Cuk - H02171-103- P011028

PAPERS AND PROCEEDINGS

OF THE

Hundred and Tenth Annual Meeting

OF THE

AMERICAN ECONOMIC ASSOCIATION

Chicago, IL, January 3-5, 1998

Program Arranged by Robert W. Fogel

Papers and Proceedings Edited by J. David Baldwin and Ronald L. Oaxaca

MAY 1998

P-11028

• Printed at Banta Company, Menasha, Wisconsin, U.S.A.

• Copyright © 1998 by the American Economic Association. Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or direct commercial advantage and that copies show this notice on the first page or initial screen of a display along with the full citation, including the name of the author. Copyrights for components of this work owned by others than AEA must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, to redistribute to lists, or to use any component of this work in other works, requires prior specific permission and/or a fee. Permissions may be requested from the American Economic Association, 2014 Broadway, Suite 305, Nashville, TN 37203.

• No responsibility for the views expressed by authors in this *Review* is assumed by the editors or the publishers, The American Economic Association.

Correspondence relating to advertising, business matters, permissions to quote, back issues, subscriptions, and changes of address, should be sent to the American Economic Association, 2014 Broadway, Suite 305, Nashville, TN 37203. Change of address notice must be received at least six (6) weeks prior to the publication month. A membership or subscription paid twice is automatically extended for an additional year unless otherwise requested.

THE AMERICAN ECONOMIC REVIEW (ISSN 0002-8282), May 1998, Vol. 88, No. 2, is published five times a year (March, May, June, September, December), and every four years it is published six times a year (twice in December), by the American Economic Association, 2014 Broadway, Suite 305, Nashville, TN 37203. Annual fees for regular membership, of which 30 percent is for a year's subscription to this journal, are: \$55.00, \$66.00, or \$77.00, depending on income. A membership also includes subscriptions to *The Journal of Economic Literature* and *The Journal of Economic Perspectives*. In countries other than the U.S.A., add \$23.00 for extra postage. Information on becoming a member can be found on the last page of this journal. Periodical postage paid at Nashville, TN and at additional mailing offices. POSTMASTER: Send address changes to the *American Economic Review*, 2014 Broadway,

Founded in 1885

Officers

President

ROBERT W. FOGEL
University of Chicago

President-elect

D. GALE JOHNSON
University of Chicago

Vice-Presidents

ROBERT J. BARRO
Harvard University
JUNE E. O'NEILL
Congressional Budget Office

Secretary

JOHN J. SIEGFRIED
Vanderbilt University

Treasurer

C. ELTON HINSHAW
Vanderbilt University

Editor of The American Economic Review

ORLEY ASHENFELTER
Princeton University

Editor of The Journal of Economic Literature

JOHN McMILLAN
University of California—San Diego

Editor of The Journal of Economic Perspectives

ALAN B. KRUEGER
Princeton University

Executive Committee

Elected Members of the Executive Committee

RONALD G. EHRENBERG
Cornell University
BARBARA L. WOLFE
University of Wisconsin
RACHEL McCULLOCH
Brandeis University
PAUL M. ROMER
Stanford University
ANGUS S. DEATON
Princeton University
LAURENCE J. KOTLIKOFF
Boston University

EX OFFICIO Members

ANNE O. KRUEGER
Stanford University
ARNOLD C. HARBERGER

THE AMERICAN ECONOMIC REVIEW

VOL. 88 NO. 2

MAY 1998

PAPERS AND PROCEEDINGS

OF THE

Hundred and Tenth Annual Meeting

OF THE

AMERICAN ECONOMIC ASSOCIATION

Chicago, IL

January 3–5, 1998

Program Arranged by Robert W. Fogel

Papers and Proceedings Edited by J. David Baldwin and Ronald L. Oaxaca

CONTENTS

Editors' Introduction	<i>J. David Baldwin and Ronald L. Oaxaca</i>	vii
Foreword	<i>Robert W. Fogel</i>	viii

PAPERS

Richard T. Ely Lecture		
Turnpikes	<i>Lionel W. McKenzie</i>	1
Clio and the Economic Organization of Science		
Common Agency Contracting and the Emergence of "Open Science" Institutions	<i>Paul A. David</i>	15
Revolution from Above: The Role of the State in Creating the German Research System, 1810–1910	<i>Timothy Lenoir</i>	22
Academic Science and Technology in the Service of Industry: MIT Creates a "Permeable" Engineering School	<i>Christophe Lécuyer</i>	28
Federal Government Initiatives and the Foundations of the Information Technology Revolution: Lessons from History	<i>Marjory S. Blumenthal</i>	34
Historical Perspectives on Current Issues of Economic Performance		
Micro Rules and Macro Outcomes: The Impact of Micro Structure on the Efficiency of Security Exchanges, London, New York, and Paris, 1800–1914	<i>Lance Davis and Larry Neal</i>	40
The Peace Dividend in Historical Perspective	<i>Hugh Rockoff</i>	46
Wages and Labor Markets Before the Civil War	<i>Robert A. Margo</i>	51
Useful Microeconomics from Business History		
Representative Firm Analysis and the Character of Competition: Glimpses from the Great Depression	<i>Daniel M. G. Raff</i>	57
Survival and Size Mobility Among the World's Largest 100 Industrial Corporations, 1912–1995	<i>Leslie Hannah</i>	62
Partnerships, Corporations, and the Theory of the Firm	<i>Naomi R. Lamoreaux</i>	66
The New Institutional Economics		
The New Institutional Economics	<i>Ronald Coase</i>	72
The Institutions of Governance	<i>Oliver E. Williamson</i>	75
Historical and Comparative Institutional Analysis	<i>Avner Greif</i>	80
Norms and Networks in Economic and Organizational Performance	<i>Victor Nee</i>	85
What We Get for Health-Care Spending		
Technological Change in Heart-Disease Treatment: Does High Tech Mean Low Value?	<i>Mark McClellan and Haruko Noguchi</i>	90
The Value of Health: 1970–1990	<i>David M. Cutler and Elizabeth Richardson</i>	97
Economic Effects of Reducing Disability	<i>Kenneth G. Manton, Eric Stallard, and Larry Corder</i>	101
Measuring Prices and Quantities of Treatment for Depression	<i>Richard G. Frank, Susan H. Busch, and Ernst R. Berndt</i>	106
Public Funds, Private Funds, and Medical Innovation: How Managed Care Affects Public Funds for Clinical Research	<i>Judith K. Hellerstein</i>	112

The Changing Market for Health Insurance

- The Demand for Medical Care: What People Pay Does Matter *Matthew J. Eichner* 117
- Adverse Selection and Adverse Retention *Daniel Altman, David M. Cutler, and Richard J. Zeckhauser* 122
- Payment Heterogeneity, Physician Practice, and Access to Care *Sherry Glied* 127
- What Has Increased Medical-Care Spending Bought? *David M. Cutler, Mark McClellan, and Joseph P. Newhouse* 132

Social Security and the Real Economy: Evidence and Policy Implications

- Social Security: Privatization and Progressivity *Laurence J. Kotlikoff, Kent A. Smetters, and Jan Walliser* 137
- Perspectives on the Social Security Crisis and Proposed Solutions *Kevin M. Murphy and Finis Welch* 142
- Social Security and the Real Economy: An Inquiry into Some Neglected Issues *Issac Ehrlich and Jian-Guo Zhong* 151

Social Security and Declining Labor-Force Participation: Here and Abroad

- Social Security and Retirement: An International Comparison *Jonathan Gruber and David A. Wise* 158
- Social Security and Labor-Force Participation in the Netherlands *Arie Kapteyn and Klaas de Vos* 164
- Pensions and Labor-Market Participation in the United Kingdom *Richard Blundell and Paul Johnson* 168
- Social Security and Declining Labor-Force Participation in Germany *Axel Börsch-Supan and Reinhold Schnabel* 173

Informing Retirement-Security Reform

- 401(k) Plans and Future Patterns of Retirement Saving *James M. Poterba, Steven F. Venti, and David A. Wise* 179
- The Cause of Wealth Dispersion at Retirement: Choice or Chance? *Steven F. Venti and David A. Wise* 185
- Socioeconomic Status and Health *James P. Smith* 192
- Extending the Consumption-Tax Treatment of Personal Retirement Saving *John B. Shoven and David A. Wise* 197

Women and Retirement Issues

- Married Women's Retirement Expectations: Do Pensions and Social Security Matter? *Marjorie Honig* 202
- Gender Differences in the Allocation of Assets in Retirement Savings Plans *Annika E. Sundén and Brian J. Surette* 207
- How Are Participants Investing Their Accounts in Participant-Directed Individual Account Pension Plans? *Leslie E. Papke* 212

Life-Cycle and Cohort Studies of Aging

- Aging in the Early 20th Century *Clayne L. Pope and Larry T. Wimmer* 217
- Rise of the Welfare State and Labor-Force Participation of Older Males: Evidence from the Pre-Social Security Era *Chulhee Lee* 222
- Secular Trends in the Determinants of Disability Benefits *Sven E. Wilson and Louis L. Nguyen* 227
- The Evolution of Retirement: Summary of a Research Project *Dora L. Costa* 232

Demographic Trends and Economic Consequences

- Uncertain Demographic Futures and Social Security Finances *Ronald Lee and Shripad Tuljapurkar* 237

Demographic Analysis of Aging and Longevity	<i>James W. Vaupel</i>	242
Aging and Inequality in Income and Health	<i>Angus S. Deaton and Christina H. Paxson</i>	248
Intergenerational Relations		
Generations and the Distribution of Economic Well-Being: A Cross-National View	<i>Timothy M. Smeeding and Dennis H. Sullivan</i>	254
Relative Cohort Size and Inequality in the United States	<i>Diane J. Macunovich</i>	259
Intergenerational Transmission of Health	<i>Dennis Ahlburg</i>	265
On the Economics of Giving		
Transfers, Empathy Formation, and Reverse Transfers	<i>Oded Stark and Ita Falk</i>	271
The Prestige Motive for Making Charitable Transfers	<i>William T. Harbaugh</i>	277
Tax Policy and Gifts	<i>Louis Kaplow</i>	283
Tax and Human-Capital Policy		
Taxes, Uncertainty, and Human Capital	<i>Kenneth L. Judd</i>	289
Tax Policy and Human-Capital Formation	<i>James J. Heckman, Lance Lochner, and Christopher Taber</i>	293
Does Government R&D Policy Mainly Benefit Scientists and Engineers?	<i>Austan Goolsbee</i>	298
Rethinking Public Education		
The Origins of State-Level Differences in the Public Provision of Higher Education: 1890–1940	<i>Claudia Goldin and Lawrence F. Katz</i>	303
How Much Does School Spending Depend on Family Income? The Historical Origins of the Current School Finance Dilemma	<i>Caroline M. Hoxby</i>	309
Demographic Change, Intergenerational Linkages, and Public Education ...	<i>James M. Poterba</i>	315
Work or Leisure: A Changing Decision?		
When We Work	<i>Daniel S. Hamermesh</i>	321
Assortative Mating by Schooling and the Work Behavior of Wives and Husbands	<i>John Pencavel</i>	326
The Unequal Work Day: A Long-Term View	<i>Dora L. Costa</i>	330
What Is Poverty and Who Are the Poor? Redefinition for the United States in the 1990's		
Absolute versus Relative Poverty	<i>James E. Foster</i>	335
Self-Reliance as a Poverty Criterion: Trends in Earnings-Capacity Poverty, 1975–1992	<i>Robert Haveman and Andrew Bershadker</i>	342
Alternative Historical Trends in Poverty	<i>David M. Betson and Jennifer L. Warlick</i>	348
Poverty-Measurement Research Using the Consumer Expenditure Survey and the Survey of Income and Program Participation	<i>Kathleen Short, Martina Shea, David Johnson, and Thesia I. Garner</i>	352
African-American Economic Gains: A Long-Term Assessment		
Race and Class in Postindustrial Employment	<i>Gerald D. Jaynes</i>	357
Quit Behavior as a Measure of Worker Opportunity: Black Workers in the Interwar Industrial North	<i>Warren Whatley and Stan Sedo</i>	363
Assessing 50 Years of African-American Economic Status, 1940–1990	<i>Marcus Alexis</i>	368
Theoretical and Empirical Developments in Cost-Benefit Analysis and Program Evaluation		
Imagined Risks and Cost-Benefit Analysis	<i>Robert A. Pollak</i>	376
General-Equilibrium Treatment Effects: A Study of Tuition Policy	<i>James J. Heckman, Lance Lochner, and Christopher Taber</i>	381

Government in Transition

- Regulatory Discretion and the Unofficial Economy *Simon Johnson, Daniel Kaufmann, and Pablo Zoido-Lobaton* 387
- Changing Incentives of the Chinese Bureaucracy *David D. Li* 393
- Private Enforcement of Public Laws: A Theory of Legal Reform *Jonathan R. Hay and Andrei Shleifer* 398

Forecasting Japan's Future: The Lessons of History

- The 1940 System: Japan under the Wartime Economy *Yukio Noguchi* 404
- Structural Change and Japanese Economic History: Will the 21st Century Be Different? *Gary R. Saxonhouse* 408
- Declining Population and Sustained Economic Growth: Can They Coexist? *Yutaka Kosai, Jun Saito, and Naohiro Yashiro* 412
- The Incentive Structure of a "Managed Market Economy": Can It Survive the Millennium? ... *Koichi Hamada* 417

China's Economic Reforms: Some Unfinished Business

- Competition, Policy Burdens, and State-Owned Enterprise Reform *Justin Yifu Lin, Fang Cai, and Zhou Li* 422
- China's State Enterprises: Public Goods, Externalities, and Coase *Gary H. Jefferson* 428
- Village Leaders and Land-Rights Formation in China *Scott Rozelle and Guo Li* 433

Banking Crises, Currency Crises, and Macroeconomic Uncertainty

- The Double Drain with a Cross-Border Twist: More on the Relationship Between Banking and Currency Crises *Victoria Miller* 439
- Financial Crises in Asia and Latin America: Then and Now *Graciela L. Kaminsky and Carmen M. Reinhart* 444
- On the Importance of the Precautionary Saving Motive *Annamaria Lusardi* 449
- Risk, Entrepreneurship, and Human-Capital Accumulation *Murat F. Iyigun and Ann L. Owen* 454

The Economics of Gun Control

- Who Owns Guns? Criminals, Victims, and the Culture of Violence *Edward L. Glaeser and Spencer Glendon* 458
- Guns, Violence, and the Efficiency of Illegal Markets *John J. Donohue III and Steven D. Levitt* 463
- Lives Saved or Lives Lost? The Effects of Concealed-Handgun Laws on Crime *Hashem Dezhbakhsh and Paul H. Rubin* 468
- Criminal Deterrence, Geographic Spillovers, and the Right to Carry Concealed Handguns *Stephen G. Bronars and John R. Lott, Jr.* 475

Teaching Statistics and Econometrics to Undergraduates

- Engaging Students in Quantitative Analysis with Short Case Examples from the Academic and Popular Press *William E. Becker* 480
- Teaching Undergraduate Econometrics: A Suggestion for Fundamental Change *Peter E. Kennedy* 487

PROCEEDINGS

- John Bates Clark Award 494
- Minutes of the Annual Meeting 495
- Minutes of the Executive Committee Meetings 496

Reports

Secretary	<i>John J. Siegfried</i>	506
Treasurer	<i>C. Elton Hinshaw</i>	509
Finance Committee	<i>C. Elton Hinshaw</i>	510
Editor, <i>American Economic Review</i>	<i>Orley Ashenfelter</i>	511
Editor, <i>Journal of Economic Literature</i>	<i>John Pencavel</i>	521
Editors, <i>Journal of Economic Perspectives</i>	<i>Alan Krueger and J. Bradford De Long</i>	524
Director, <i>Job Openings for Economists</i>	<i>C. Elton Hinshaw</i>	526
Committee on Economic Education	<i>Michael K. Salemi</i>	528
Committee on the Status of Minority Groups in the Economics Profession ...	<i>Susan M. Collins</i>	529
Committee on the Status of Women in the Economics Profession	<i>Robin L. Bartlett</i>	532
Representative to the Social Science Research Council	<i>Michelle J. White</i>	536
Representative to the National Bureau of Economic Research	<i>John J. Siegfried</i>	537

THE purpose of the American Economic Association, according to its charter, is the encouragement of economic research, the issue of publications on economic subjects, and the encouragement of perfect freedom of economic discussion. The Association as such takes no partisan attitude, nor does it commit its members to any position on practical economic questions. It is the organ of no party, sect, or institution. People of all shades of economic opinion are found among its members, and widely different issues are given a hearing in its annual meetings and through its publications. The Association, therefore, assumes no responsibility for the opinions expressed by those who participate in its meetings. Moreover, the papers presented are the personal opinions of the authors and do not commit the organizations or institutions with which they are associated.

Editors' Introduction

This volume contains the *Papers and Proceedings* of the one-hundred and tenth annual meeting of the American Economic Association. The *Proceedings* record the business activities of the Association in 1997, the annual membership meetings, and the March and January (1998) meetings of the Association's officers and committees. The *Papers* constitute the greater part of the volume. They comprise contributions that fill roughly the same number of pages as two regular issues of *The American Economic Review*. We will take this opportunity to answer a number of commonly asked questions about the *Papers*.

Who chooses the authors? About a year in advance, the Association's President-elect, acting as program chairman, decides on the topics for which sessions will be organized. This is done after consultation and comment, both volunteered and solicited, from a wide range of individuals. (A *Call for Papers* appears in the Summer and Fall issues of *The Journal of Economic Perspectives*.)

The President-elect invites some sessions and selects additional sessions from the various proposals received. Each session organizer in turn invites several persons (usually three or four) to give papers on the theme of the session and asks others to give comments on the papers. The program chairman decides at the time of organization which sessions will be included in this volume. Space limitations restrict the number of printed sessions. This year we are printing 27 sessions, although a total of 151 sessions were sponsored, either solely by the American Economic Association or jointly with other allied societies.

Are discussants' comments published? Comments and discussions are not published. For all sessions, names and affiliations of commentators are printed at the start of each session, permitting readers especially interested in particular comments to write to the commentator for a copy of the discussion.

What standards must the papers meet? The *American Economic Review's* policy regarding availability of data also applies to the *Paper and Proceedings*: papers are published only if the data used in the analysis are clearly and precisely documented and are readily available to any re-

searcher for purposes of replication. Otherwise, the guidelines under which papers are published in the *Papers and Proceedings* differ considerably from those governing regular issues of the *Review*. First, the length of papers is strictly controlled. They must be no more than 12 typescript pages in three-paper sessions, and 10 typescript pages in four-paper sessions. Second, papers are not subjected to a formal refereeing process. However, a paper can be rejected if, after reading it, we conclude that it is utterly without merit. Third, the content and range of subject matter reflect the wishes of the President-elect to investigate and expose the current state of economic research and thinking. In most cases, therefore, the papers are exploratory and discursive, rather than formal presentations of original research.

In order to produce this volume by May, strict deadlines must be met, and there is no time for communication with every author about editing changes made in order to improve content and style and to satisfy space restrictions. Every effort is made to notify an author prior to the deadline if the paper is too long or does not satisfy other specifications.

For the most part, authors were quite cooperative this year, and for this we are grateful. We thank them for making our job easier.

Acknowledgments. The extremely tight production schedule of this issue requires a highly coordinated chain of events; every link in that chain must be a strong one. Especially this year, we are indebted to a group of highly talented individuals for their help in bringing this project to fruition: Kathy Simkanich and Irene Rowe in the Princeton *AER* office capably handled the voluminous correspondence associated with this issue, and Laurie Burton served as proofreader extraordinaire. The staff of CJS-Tapsco, our typesetters met the challenge of keeping the issue on schedule, and we are especially grateful to Barbara Stabb for overseeing the typesetting with diligence and good humor. Finally, as in past years, we thank Kathy Holewinski and the Banta Company for their dependable work in printing and distributing this issue.

J. DAVID BALDWIN
RONALD L. OAXACA

Foreword

Because the American Economic Association has such a large and vigorous membership, more worthy papers are proposed each year than can be accommodated at the annual meetings. This year was no exception. The main burden of sorting through the proposals was borne by the members of the Program Committee, which included John Cochrane, David M. Cutler, Ronald A. Dye, Richard B. Freeman, Bruce A. Grundy, Oliver Hart, James J. Heckman, John Kain, Alan Kelley, Glenn Loury, Deirdre McCloskey, Rachel McCulloch, William Nordhaus, Sam Peltzman, James M. Poterba, Paul Romer, Andrei Shleifer, Peter Temin, and David A. Wise.

The President-elect, who is a novice in the organization of such a large set of meetings, drew heavily on the advice and aid of the highly professional AEA staff. C. Elton Hinshaw, John Siegfried, and J. David Baldwin instructed me on my duties and passed on the experience of the past. Violet Sikes and Marlene Hight answered numerous urgent queries and smoothed the way. Karen L. Brobst, who served as my administrative assistant for the meetings, actually bore the brunt of the work at Chicago. I benefited from the advice and experience of Arnold C. Harberger and Anne O. Krueger, the two immediate past-presidents, and from Enid M. Fogel.

I did not have an overall theme for the Chicago meetings but singled out three areas for the invited sessions that I believe are especially important for the economics profession, both currently and in years ahead. These themes are the economics of aging, including pensions, health care, and leisure; continuing and new issues of economic equity; and the rise to dominance of a set of emerging market economies with special emphasis on Asia. There is an implicit fourth category of invited sessions that are of high quality in a variety of areas of economics. No effort was made to restrict the contributed papers to these areas,

since I felt that the contributed papers should reflect the full range of research underway among economists.

One-third of the invited sessions published here deal with problems in the economics of aging, broadly defined. Not only the United States but virtually all OECD nations are faced with a crisis in their pension and health-care systems, not because they are poor, but because they are, by historical or third-world standards, exceedingly rich. It is the enormous increase in their per capita incomes over the past century that has permitted the average length of retirement to increase by fivefold and the amount of leisure time available to those still in the labor force to increase by nearly fourfold. The increased spending on health care, pensions, and leisure is a concomitant of the synergism between technological and physiological improvements that has produced a form of human evolution that is biological but not genetic, rapid, and culturally transmitted but not necessarily stable. This process, "technophysio evolution," is still ongoing in both rich and developing countries. The papers in this issue underscore the need for basing economic policies with respect to aging on longitudinal studies. Since short-term income elasticities of demand for retirement and health care may differ considerably from the corresponding long-term elasticities, it is necessary to take account of the evolutionary processes that influence the long-term elasticities.

Collectively, the papers on aging point to the need to expand and refine prevailing theories on the economics of leisure. A century ago, hardly one-sixth of lifetime discretionary hours was spent on leisure-time activities. Today more than half of lifetime discretionary hours is devoted to leisure. Curiously, leisure remains a stepchild in theoretical work and, despite a number of efforts to correct the practice, is still omitted from national income accounts.

ROBERT W. FOGEL

RICHARD T. ELY LECTURE

Turnpikes

By LIONEL W. MCKENZIE*

I will sketch the history of the so-called turnpike theorems and describe some interesting recent developments regarding them. I will also discuss the attempt to apply the ideas of turnpike theorems in the literature of optimal capital accumulation to the theory of competitive equilibrium over time. Finally I will make some remarks on the relation of this literature to some recent developments in the theory of economic growth which are often referred to comprehensively as the New Growth Theory. Of course the New Growth Theory like the Old Growth Theory is not concerned directly with optimal capital accumulation, but with the actual course of events in markets, which indeed need not be perfectly competitive. I do not pretend for a moment to have a comprehensive mastery of this enormous literature. Especially I must apologize to those many able economists, beginning with William A. Brock and Leonard J. Mirman (1972), who have developed turnpike theorems under uncertainty. Time does not permit my discussing their work, nor does my command of the literature. Finally, I have tried to pitch my lecture at a level that does not require a great deal of prior knowledge of the subject.

I. Origins

I do claim to have one qualification for delivering the Ely lecture. I learned my first economics from Ely's textbook in a small junior college in middle Georgia, named appropri-

ately Middle Georgia College. I should add that the summer after sitting in the Principles class I did consult an even earlier and more highly revered source, *The Wealth of Nations*, read from the Harvard Classics. Somehow this volume prevailed over *The Origin of Species* from the same collection and led me to economics rather than biology, although my scholarship to Duke was obtained with the aid of my biology professor. He never responded to my letter informing him of my apostasy.

There are two principal sources of the modern theory of optimal capital accumulation as well as modern growth theory. They were very nearly simultaneous. The earlier to be published was Frank Ramsey's (1928) "Mathematical Theory of Saving," one of Ramsey's three great contributions to economics, this one on the suggestion of John Maynard Keynes. The other is the equally famous paper by John von Neumann which was first delivered to a mathematics seminar in Princeton in 1932. This was one of his two great contributions to economics. It was also given in 1936 to a mathematics colloquium in Vienna led by Karl Menger, the son of Carl Menger, the economist. I heard the paper in an economics seminar in Princeton around 1940. I can provide partial confirmation of Oskar Morgenstern's remark that no one in the audience understood a word of it. It was translated and published in the *Review of Economic Studies* in 1945 under the title "A Model of General Economic Equilibrium." It is rather intriguing that both Ramsey and von Neumann were mathematicians of the first rank. I do not believe the loss of any other five papers from the economics literature would have had a greater impact on the development of economic theory.

Ramsey assumes there is one good which serves both for capital accumulation and for consumption. The good is produced by capital

* Department of Economics, University of Rochester, Rochester, NY 14627. This academic year marks the 40th anniversary of the founding of the Rochester Economics Department. Therefore, I wish to dedicate this paper, unworthy though it is, to the faculty and graduate students of the Rochester Department over those years. Many of their names will appear in my paper. Even more should appear were there more time to discuss their contributions.

and labor. In addition there is a social utility function with this good and labor as arguments. An essential assumption for his principal result is that the economy can achieve, or at least asymptotically approach, satiation, either in production or utility, a condition he called "Bliss." Then without assuming that future utility is discounted he derives a rule for capital accumulation which realizes the maximum sum of utility over time. In those days in Cambridge, England, discounting future utility over generations was not favored. Ramsey's famous criterion was, and I quote, "[the] rate of saving multiplied by [the] marginal utility of consumption should always equal bliss minus [the] actual rate of utility enjoyed" (Ramsey, 1928 pp. 547-49). Keynes gave him a nice intuitive way of seeing that this criterion is correct. However, the rule which he also derived and which remains central to the modern optimal growth theory is that in the absence of discounting, the marginal utility of consumption should fall at a proportionate rate given by the rate of interest. In perfectly competitive markets this is the marginal product of a unit of capital.

The contribution of von Neumann was along very different lines. Ramsey uses what today is called a macro model; von Neumann uses a disaggregated general-equilibrium model, whose production sector is an activities model. It has the peculiarity that there is no explicit recognition of labor inputs. Capital goods produce capital goods. Moreover there is no utility function in his model. He sought to show that there is an economic equilibrium in the sense that prices exist which will allow activities in use to cover costs while no activity offers a positive profit. Moreover, the activities in use can expand the stock of capital goods at a maximal feasible rate. He proved that the rate of expansion of the capital stock and the rate of decline in prices will be equal in this equilibrium. He also generalized a famous fixed-point theorem due to L. J. Brouwer which later played a critical role in proving the existence of a competitive equilibrium.

When von Neumann presented this result in Cambridge, Massachusetts, he asserted that maximization of an objective function had no part in his theory. Paul Samuelson who was in the audience rose to challenge this statement,

asserting that maximization would enter once disequilibria were considered. Von Neumann offered to bet him a cigar that this was wrong. Amazingly Samuelson did not accept the wager. Nonetheless he feels that should they meet at St. Peter's gate he should ask von Neumann for a cigar. It was in the context of a von Neumann model, in a publication of the Rand Corporation, that Samuelson (1949) first described the idea of the turnpike.

II. The Samuelson Turnpike

The subsequent history of models of optimal growth has featured an interplay of these two foundations. That is, the Ramsey objective of maximizing a utility sum over time has been introduced into the disaggregated model of von Neumann, and the von Neumann production sector featuring numerous activities has been introduced into the Ramsey model. It is the accomplishment of the von Neumann model to describe the conditions for an equilibrium of production over time.

Samuelson conjectured the existence of a turnpike in a von Neumann model where the objective to be maximized is the size of the terminal capital stocks in certain assigned ratios. The turnpike was to be the path of most rapid balanced growth of the capital stock, the von Neumann equilibrium. Therefore, it seems appropriate to refer to the turnpike in the von Neumann model as the Samuelson turnpike. Later Robert Dorfman, Samuelson, and Robert Solow (1958) presented a proof of the Samuelson turnpike conjecture for a model with two capital goods. However, the first rigorous proof was found by Roy Radner (1961). His model allowed any number of capital goods. He also introduced the most useful method of proof for turnpike theorems, which I have called the value-loss method.

The value-loss method exploits the fact that, compared with the capital stocks of paths on the turnpike, the capital stocks of paths off the turnpike lose value at the equilibrium prices that support the turnpike. Since it is possible to use the turnpike for most of the time in an alternative path, the losses that the optimal path can suffer are limited. This limits the time the optimal path can spend outside a neighborhood of the turnpike, just as for your trips

by car through the countryside. Radner showed that any sufficiently long optimal path in an irreducible model will spend most of the time in a small angular neighborhood of the ray on which von Neumann equilibrium paths lie. He assumed that the production set, which is a convex cone, is strictly convex near the von Neumann ray except for constant returns to scale. This theorem has the weakness that strict convexity is not consistent with the neoclassical production model, for which different industries have independent production processes. The difficulty is that the social production set is convex but not strictly convex under neoclassical assumptions. However, the argument can be adapted to prove convergence to a flat piece of the production set on which the von Neumann equilibrium path lies, which I call the von Neumann facet. More of that later. Other arguments on some additional assumptions imply that the optimal path will converge further to the von Neumann equilibrium itself.

Using approaches different from Radner's, Michio Morishima (1961) and I (McKenzie, 1963a) independently proved a turnpike theorem for a von Neumann model of Leontief type with circulating capital and variable coefficients. As a von Neumann model there are no explicit labor inputs. This model has a neoclassical production sector. My approach was to use a theorem of Solow and Samuelson (1953) which implies that the equilibrium prices converge to the turnpike prices over time. This causes the production coefficients to converge to the coefficients that produce the turnpike. Then, tracing the production path backwards in time I show that the optimal path must stay near the turnpike most of the time if the period of accumulation is long enough. It has recently been shown by Michael Kaganovich (1998) that it is possible to introduce a utility function into this model and prove a turnpike theorem where the objective is to maximize a discounted utility sum over the infinite future. More about this later.

III. The Ramsey Turnpike

The first moves beyond the theorem of Ramsey, with its assumption of one sector and the state of Bliss as the turnpike, were made

independently by Tjalling Koopmans (1965) and David Cass (1966). In the modern literature, the state of the economy where population is allowed to grow and saturation has been reached in the sense of a maximum of sustainable per capita utility is sometimes referred to as a Golden Age. If discounting is allowed as well, an optimal path of accumulation with constant per capita capital stock is called a modified Golden Age. Cass and Koopmans gave rigorous proofs of convergence of optimal paths to a modified Golden Age where utility is discounted and population is growing. Koopmans used a social utility function defined on per capita consumption plus possible additional discounting, while Cass equivalently assumed that the discount rate on social utility exceeds the rate of population growth. Ramsey had introduced generalizations in these same directions, but his arguments were not understood and probably not complete. The theorems of Cass and Koopmans were subsequently generalized by James Mirrlees (1967) to allow technical progress as well.

Some essays in the New Growth Theory take their departure from a one-sector model of this type, that is, the Ramsey model into which population growth, technical progress, and discounting have been introduced with the rate of discounting of social utility exceeding the sum of the rate of population growth and the rate at which individual utility increases because of technical progress. In such a model, the New Growth theorists seek to prove turnpike theorems in the sense of convergence of the capital stock to the capital stock of a modified Golden Age. This literature is comprehensively surveyed in the recent book *Economic Growth* by Robert Barro and Xavier Sala-i-Martin (1995).

One should note that the meaning of turnpike in the one-sector model of Ramsey is just the level of capital accumulation reached at a saturation point, either a Golden Age or a modified Golden Age. On the other hand, in the von Neumann model the turnpike is given by the combination of capital goods that attains the fastest growth rate. When this idea is adapted to a Ramsey model where there are resources like labor and land that cause production to be bounded, the comparable task

is to discover the combination of capital goods that supports production in the state of saturation. This does not exclude sustained growth so long as it is made possible by exogenous factors like population growth and technical progress. When there is more than one capital good, the combination of capital goods that can play the rôle of the turnpike must be found, that is, to which optimal paths converge.

The first proof of a turnpike for a Ramsey model with more than one sector was found by Hiroshi Atsumi (1965) in a model with one capital good and one consumer good and no discounting. He assumed, like Koopmans and Cass, an expanding population, and he used the Ramsey objective based on undiscounted per capita utility. He introduced the overtaking criterion so that this objective would be meaningful over an infinite future. He also relates his argument to the competitive market and derives the optimal savings ratio in the sense of Ramsey. His argument is a value-loss argument, and he derives a generalization of the lemma that was the basis of Radner's proof of the Samuelson turnpike theorem. However, Atsumi's theorem does not achieve full generality since he uses only one capital good.

The turnpike theorem for any number of capital goods and consumers was made in a model with a finite horizon by me (McKenzie, 1968) and in a model with an infinite horizon by David Gale (1967). Gale assumed strict concavity of social utility, at least at the Golden Age, while I considered the case where the concavity was not strict. Utility was not explicitly discounted, but for an expanding population an interpretation of our models in terms of per capita utility would implicitly amount to discounting social utility at the rate of population growth. Of course the general model should allow for greater levels of discounting. However, it took a surprisingly long time to prove turnpike theorems with greater levels of discounting. The extension to a multisector model with discounting exceeding the rate of population growth was made independently by Cass and Carl Shell (1976) and José Scheinkman (1976). They proved the convergence of optimal paths to the modified Golden Rule path. Their theorems, like the theorem of Gale (1967), require strict concavity of the

reduced utility function. The reduced utility is the maximum social utility achievable in a single period, given the initial and terminal capital stocks. Concavity of the reduced utility function is implied by strict concavity of the social utility function, defined on consumption, and strict convexity of the production possibility set, except for constant returns to scale. The theorems are stated in terms of discount factors, net of population growth, which are close enough to 1, that is, net discount rates sufficiently close to 0, so that the optimal path from certain initial capital stocks converges to an optimal balanced path. The optimal balanced path itself converges to the unique optimal balanced path for the undiscounted case as the discount rate goes to 0. Thus, one way of interpreting their theorems is that the optimal paths are continuous in the discount factor as the discount factor approaches 1.

IV. Neighborhood Turnpikes and von Neumann Facets

Given strictly concave utility and strictly convex production sets, convergence to an optimal balanced path, or a modified Golden Age, may require that discount rates be very close to 0, or equivalently discount factors close to 1. However, I have shown (McKenzie, 1986) that this requirement may be relaxed if convergence to a modified Golden Age is replaced by convergence to a neighborhood of a modified Golden Age. Then the discount rate must be closer to 0 the smaller the neighborhood chosen. The modified Golden Ages need not be unique, but they will all lie in similar neighborhoods of each other. The differentiability requirements on the utility and production functions are weaker in my theorem. Later theorists have found that all kinds of complex patterns of optimal paths are possible even when differentiability is assumed. For example, periodic paths or even chaotic paths may occur. Such cases have been extensively investigated and described by Michele Boldrin, Tapan Mitra, Kazuo Nishimura, Makoto Yano, and many others. Much of the literature is surveyed by Boldrin and Michael Woodford (1990). However these complex patterns must lie within the neighborhoods employed by the neighborhood

turnpike theorems, when the reduced utility function is strictly concave at the modified Golden Age. These neighborhoods close down on the modified Golden Age as the discount rate approaches zero.

The added generality I will now describe when strict convexity and concavity assumptions are relaxed cannot be too easy to grasp since Tjalling Koopmans told us at Stanford in 1965 that he did not understand it. However, Roy Radner was present and said that he did understand it, which somewhat relieved my mind. I first introduced (McKenzie, 1963b) this generalization with reference to Radner's theorem on the Samuelson turnpike where Radner assumed strict convexity of the production possibility set at the von Neumann ray except for constant returns to scale. However, I will confine attention here to the Ramsey model. In the absence of strict concavity of the reduced utility function, the prices that support the modified Golden Age may also support a convex set of input-output combinations which surround the modified Golden Age, perhaps including many which are not balanced. In this case, what the value-loss argument produces is not convergence to the modified Golden Age, but to the set of input-output combinations for capital stocks that are supported by the same prices that support the modified Golden Age. I call this convex set of input-output combinations a von Neumann facet in the Ramsey model. The convergence theorem requires the supporting price vector to be unique and the value losses off the facets to be uniformly bounded from zero (McKenzie, 1983). The theorem leaves open how the optimal path will behave within the neighborhood, even when the von Neumann facet is trivial and contains only the modified Golden Age. Complex patterns may also arise on the von Neumann facet, when it is not trivial. In this case they may be cyclic but not chaotic. On the other hand, it is also possible that the difference equations that govern paths which lie on the facet will be such that convergence to a small neighborhood of the Golden Age must also occur for a small enough discount rate. These considerations are rather too complicated to describe more exactly in this lecture. Unfortunately, however, the secondary

sources are very inadequate on neighborhood convergence and on von Neumann facets as well.

The neoclassical model always has non-trivial von Neumann facets. Given the prices of inputs and outputs in each production process, the most profitable combination of inputs and outputs will be chosen, and with perfect competition the profit will be zero. Through the replication of the elementary production units, assuming they are small, the level at which these combinations are realized may be varied nearly continuously. The consequence is that a great variety of total inputs and outputs are consistent with given prices. Moreover, although the dimension of this variation of inputs and outputs is reduced if the variation in activity levels is constrained by side conditions, it is not eliminated so long as the side conditions are fewer in number than the number of processes. For example, the total supplies of the primary resources may be side constraints. These relations have been discussed in a model without joint production by Harutaki Takahashi (1985).

In the model with bounded paths and differentiability of social utility near the von Neumann facet, Makoto Yano (1998) has recently proved a dual turnpike theorem, which is stated in terms of prices rather than goods, on assumptions significantly weaker than those used for the primal theorem, which is stated in terms of goods. My theorem requires uniformity in value losses off the von Neumann facets for an interval of discount rates bounded by 0. His theorem does not require this uniformity. The dual turnpike theorem asserts that the prices that support the optimal path will converge to a neighborhood of the unique price vector that supports a modified Golden Age. In a competitive economy where perfect-foresight paths are optimal paths his theorem implies that fiscal policies that transfer income from some consumers to others will not affect the spending of these consumers to any significant extent if the transfers are temporary. This extends results of Milton Friedman (1957) based on the permanent-income hypothesis from a partial-equilibrium framework to a general-equilibrium framework.

V. Unbalanced Growth with Bounded Paths

Almost all the attention to asymptotic convergence has been concentrated on convergence to balanced paths, although it is not clear that optimal balanced paths will exist. This type of path is virtually impossible to believe in, if the model is disaggregated beyond the division into human capital and physical capital, and new goods and new methods of production appear from time to time. However, the fundamental convergence results in models in which paths are bounded, after allowing for population growth and exogenous technical progress, do not depend on the presence of an optimal stationary path or the absence of changes in technology and taste. The basic result is that optimal paths from different initial stocks have a tendency to converge, whatever their shapes may be. I made this point in articles published in 1974 and 1976. It was also emphasized in my chapter in the *Handbook of Mathematical Economics* (1986). If the discount factor is equal to 1, that is to say, future utilities are treated on a par with the utility of the current period as Ramsey would prefer, and the reduced utility function is strictly concave, it is a general phenomenon that optimal paths from different starting points converge in some sense in models with bounded growth if the paths are not isolated—in other words, if the optimal path from one starting point can be reached from the other starting point. Of course, if the optimal path from the second starting point can be reached only with great difficulty from the first starting point, the convergence may be slow.

There is a simple argument which does not use value losses that proves this result. We are free to normalize utility so that the utility along one optimal path is zero in every period. Also assume that the starting point of this path is interior to the set of capital stocks whose utility sums are bounded above minus infinity after the normalization is made. Suppose the optimal path from the second starting point stays away from the optimal path from the first starting point by at least a small distance for an indefinite number of periods. Now consider a path lying halfway between these paths. By convexity of the production set, it is feasible. If this convexity is uniform, which is not un-

reasonable if the path is bounded, the midpath will enjoy a utility that exceeds the average utility of the two paths by more than some positive amount in every period when they are apart by more than a given distance. If there are an indefinite number of such periods, and no discounting, the total utility along the midpath will exceed the average of the total utilities along the two original paths by an unbounded amount. In other words, the utility gain is infinite over the infinite path. I cannot give a complete argument since I do not wish to use even simple algebra in this lecture, but it is easily shown that this leads to a contradiction. The contradiction can only be escaped if the paths converge.

If the discount factor is less than 1, the obvious difficulty arises that the utility gains of the midpath will not be unbounded, but will have a finite sum. However, an argument adapted from Truman Bewley (1982) still allows a neighborhood theorem to be proved. Consider the gains of the midpath over the average of the two optimal paths from an arbitrary time from which the path is observed. It can be shown, if the discount factor is close enough to 1, that the sum of future gains decreases over time by at least a certain amount if the path from the second starting point remains outside a given neighborhood of the path from the first starting point. This argument, as the earlier one, depends on uniform concavity of the reduced utility function, at least in the neighborhood of one of the paths. Since the sum of gains cannot become negative, we reach a contradiction unless the paths converge. This theorem can be given a generalization to facets if strict concavity fails to hold.

VI. The Old Growth Theory

The work that I have been reviewing should be called the Old Optimal-Growth Theory. Solow (1956) and Trevor Swan (1956) introduced what is properly called the Old Growth Theory. They are concerned with competitive equilibria. The crucial assumption in Solow's model is that saving by consumers is fixed at a percentage of net output or net income. Production is guided by current prices as in a Walrasian model with stationary expectations.

With these behavioral assumptions and strongly diminishing returns to capital, given the labor supply, Solow shows that the competitive equilibrium over time will approach a stationary value, that is, a state in which saving is just adequate to meet the demand for capital arising from the expanding population. If population is not growing and saving is a proportion of net output, capital will continue to grow. On the other hand, it seems unlikely that people would continue to accumulate capital after additions to capital no longer increase, or may even reduce, output. Thus it is necessary to introduce some dependence of the saving rate on the level of accumulated capital. This dependence presumably requires some attention to utility considerations and some foresight. If there is exogenous technical progress which takes the form of the increasing productivity of labor, the analysis may be carried through using units of effective labor rather than units of labor. The Old Growth Theory was introduced to refute the theories of Roy Harrod and Evsey Domar, which used fixed coefficients of production and implied that equilibria would always be unstable. It accomplished this purpose by bringing economic considerations into the choice of inputs for production, but it left consumption levels still divorced from economizing choice over time.

Solow has a very clear presentation of the view of the old growth theorists toward the positive and normative theories in his Radcliffe lectures (1970). There is no reluctance to discuss the question of what saving rate would be optimal, and the analysis is conducted in the same style that Ramsey used. However, the normative theory is placed squarely in the realm of policy, and there is no suggestion that the competitive market by itself will achieve a normative result. On the other hand, it is suggested that a loose approximation to the optimal policy may be good enough.

There is a model of population growth that bears a suggestive similarity to the von Neumann model and the Old Growth Theory in its positive form. The Old Growth Theory assumes a rate of saving and an aggregate initial stock of capital and explores the implications in a model of capital accumulation. Turnpike theorems in the von Neumann model

assume an initial stock of capital goods of arbitrary composition and follow the evolution of the stock where all output in excess of workers' subsistence is reinvested, and production processes satisfy no profit conditions. The demographic growth theory assumes birth rates and death rates by cohort and an initial composition of the population and traces the subsequent development of the population by size and age distribution which these assumptions imply. If the birth and death rates are assumed to be constant, the composition of the population is shown to converge to a constant composition which is independent of the initial composition, and the rate of growth of the population becomes constant as well. This development does not correspond exactly to either of the economic models but there are analogies to both the von Neumann and the Old Growth Theory. However, the demographic analogy goes further. If the birth and death rates are assumed to change over time, it still holds true that the composition of the population by age is asymptotically independent of the initial composition of the population. We found for the Ramsey model of bounded growth that the path itself was asymptotically independent of the initial capital stocks. However, convergence in the demographic model will be like convergence in the von Neumann model in that the size of the later populations will be proportionate to the initial population for a given initial composition. The growth rates will converge in both cases. This demographic theory was first conjectured by Ansley Coale (1957) and successfully used by him to make population projections. The theorems were proved by Coale's student Alvaro Lopez (1961).

We may observe that the demographic theory is subject to the same exception that has been taken to the Old Growth Theory. That is, the decisions of the population about the size of families are not explained by utility-maximizing choices made in face of anticipated incomes and death rates but are assumed to remain at certain levels, probably those realized in the past. There have been efforts by economists, in particular by Gary Becker et al. (1990), to introduce economizing into demographic theory. There is the same dispute among demographers as among economists on

whether these moves toward optimizing models have been useful.

VII. Optimal Growth and Competitive Equilibrium

So far I have discussed turnpikes chiefly in terms of optimal growth, which is a normative theory. However, there have been applications of the optimal growth theory to the market economy. Robert Becker (1982) showed how the Ramsey description of optimal growth could be placed in correspondence with Irving Fisher's theory of markets over time. Bewley (1982) showed, using the methods that Takashi Negishi applied to proving the existence of competitive equilibrium, that if perfect competition and perfect foresight are assumed then paths over time are optimal paths for a social welfare function which is the sum of individual utilities, weighted in a certain manner. The weights are the reciprocals of the marginal utilities of wealth for the individual consumers, that is, of whatever is used to state prices. These do not represent moral values. The theory is positive, not normative. The consumers make decisions that allow for bequests, so that their decisions represent consumption plans into the indefinite future.

A further step is to show that competitive paths over the infinite horizon will have a turnpike property as a consequence of being optimal paths given the appropriate welfare function. However, some caution is needed when a competitive path is characterized as a turnpike. The social utility function which is maximized along the dynamic equilibrium path depends on the initial conditions. This is because the weights given consumers in the equilibrium are dependent on the income distribution and therefore depend on the distribution of ownership in the initial stocks and on the level of these stocks. However, Yano (1984) shows that the effect of these initial conditions tends to disappear as the discount rate on utility approaches zero. Then we obtain a turnpike theorem in the original sense.

Most of the applications of turnpike theory to competitive equilibrium have been made under quite restrictive assumptions. First it is assumed that individual utility over time can

be represented by utility functions that are separable and additive over time and discounted at a constant rate. In a period model with long periods the assumption of additivity and separability may not be very demanding, though certainly not without objection. However, it is much less acceptable to assume that all consumers have the same discount rates on utility. If discount rates are allowed to be endogenous and dependent on levels of wealth, the assumption of equal discount rates may be somewhat more acceptable. This move was introduced by Hirofumi Uzawa (1969) for an aggregative model. More recently recursive utility functions for individual consumers have been introduced by Robert E. Lucas and Nancy L. Stokey (1984) who prove turnpike theorems when, loosely speaking, the discount rates on utility increase with increasing wealth. This discourages runaway saving by rich people. Recursive utility is treated in considerable detail by Robert Becker and John H. Boyd III (1997) in their book *Capital Theory, Equilibrium Analysis and Recursive Utility*.

VIII. The New Growth Theory

At this point we are left with two quite different turnpike theories. One theory is concerned with economies that expand indefinitely at rates determined endogenously. When the objective is to maximize terminal stocks the economy eventually expands at the maximal possible rate. The other theory is concerned with economies that may expand indefinitely but at rates determined outside the economic model, perhaps by the rate of growth of population. Moreover by adding the population growth rate to the discount rate and using this as an expanded discount rate, the economy can be viewed as moving to a saturation level for discounted utility, a modified version of Ramsey's Bliss. The question then arises whether, apart from population growth, the economy can grow indefinitely from endogenous factors in the manner depicted by the von Neumann growth model. And if this may occur does this economy have turnpike properties as the earlier ones do.

The first paper in the optimal-growth literature in which the rate of technical progress was made endogenous seems to be that by

Uzawa (1965) in a model with one produced good serving both as capital good and as consumption good. He assumed that technical progress depends on the portion of labor devoted to the educational sector which leads to increased labor productivity in the form of Harrod neutral technical progress. His paper has a normative slant. This type of model was developed further by Lucas (1988) 23 years later with more positive intentions. The interest of Uzawa's model was somewhat reduced by his assumption that utility is proportional to consumption. Lucas uses a constant-elasticity consumption function and a Cobb-Douglas production function with Harrod neutral technical progress which he describes as resulting from accumulating human capital. Also, he adds a term to represent external effects from the accumulation of human capital in the economy. As a consequence of this factor, competitive equilibria do not realize Pareto optima. It should be noted, however, that sustained growth in his model does not depend on the presence of the external effects.

On the other hand, somewhat earlier Paul Romer (1986) used external economies in the production of goods arising from the spread of innovations to make possible indefinite expansion. This occurs in the context of competitive equilibrium, but because of the external economies the equilibrium is not Pareto optimal. In a later model proposed by Romer (1990) knowledge is produced by research using human capital as an input to design new intermediate products. The quantity of human capital is constant and separate from the labor supply. The human capital is divided between research and production.

The papers of Romer and Lucas began an avalanche of papers in this style, which continues unabated. Uzawa proved convergence of the optimal path to a balanced path, that is, a turnpike. The problem is simplified by the fact that only one good is present so that it is not necessary to choose the combination of capital goods that appear on the optimal balanced path. Romer and Lucas were primarily interested in the path under free enterprise, and they conjectured convergence to balanced paths in these economies but did not pursue the question.

Note that the New Growth Theory is distinguished by two principal characteristics. There is continuing growth at endogenously determined rates, and the growth path is treated as a general equilibrium of a competitive market, but not necessarily a perfectly competitive market. The use of general equilibria of imperfectly competitive markets in growth theory was pioneered by Romer (1986).

Subsequently there have been many efforts toward proving turnpike theories in models of the New Growth Theory, sometimes with two capital goods denoted physical capital and human capital. As in the cases of Uzawa, Lucas, and Romer, these models have allowed for unbounded growth even without population growth. If Harrod neutral technical progress is possible in the production sector from devoting a part of labor to the education sector, the possibility of continued growth is clear. Of course it is just the same if human capital can be produced by the use of human capital alone. Indeed it is also possible if the production of human capital requires both goods and human capital so long as neither goods production nor the production of human capital is constrained by a given labor supply. The conditions that are necessary for such continued growth were recently described by Jim Dolmas (1996). What is needed is that some subset of capital goods can be produced out of themselves. This subset of capital goods is analogous to the set of basic goods used by Piero Sraffa in his book *Production of Commodities by Means of Commodities* (1960).

If such a subset of basic goods exists, unlimited expansion in the supply of other goods can occur even if their production does involve fixed factors, provided there is sufficient substitutability between the basic goods and the fixed factors. For example Cobb-Douglas production functions suffice. Let labor and capital be the two factors in a production function for a consumption good and assume that constant returns prevail, but allow capital to reproduce itself. Then the rate at which the production of the consumption good increases along a balanced path is equal to the product of the share of capital in the output of the consumption good times the own rate of return of capital which is determined in the industry producing the capital good. If the share of capital in the

consumption goods industry is constant, the rate of expansion in the supply of the consumption good is constant though less than the rate of expansion of the supply of capital. The return to capital per unit in terms of consumption goods approaches zero although it does not reach zero. If the utility function depends only on consumption and has a constant elasticity of substitution, there will be an optimal balanced path which is optimal from the appropriate initial stocks provided the discount rate is sufficiently high, so the utility sum is finite. The conditions for the existence of an optimal balanced path are expounded in the general case by Dolmas (1996). They were first stated for a one-sector model in an early paper by Brock and Gale (1969).

IX. Turnpikes and Endogenous Growth

As I mentioned earlier Michael Kaganovich (1998) has shown that my method of proving the Samuelson turnpike theorem in a model of production without joint production can be extended to a Ramsey turnpike theorem where the objective is the sum of discounted future utility. The Ramsey objective first was introduced into a Leontief model with fixed coefficients by Atsumi (1969). Kaganovich introduced variable coefficients along the lines of my proof of the Samuelson turnpike theorem. But he goes further to prove that the linearity of the model only need hold asymptotically.

One of the models of the New Growth Theory is a special case of Kaganovich's model. It is a feature of the von Neumann model that growth is sustained, and the rate of growth is endogenous to the model. Since sustained growth is the principal distinguishing feature of the New Growth Theory it is not surprising that the model of von Neumann and the Samuelson turnpike theorem should prove relevant. However, the economic significance of von Neumann's model has been questioned from the fact that it does not recognize the supply of labor as a constraint on production. The New Growth Theory in one of its incarnations tries to avoid this criticism by replacing raw labor by human capital which is assumed to be reproducible without limit out of human capital and perhaps physical capital. The

model in the New Growth Theory has the utility of consumption as the objective, but this is not a barrier to balanced growth if the utility function is homogeneous of some degree, a usual assumption the New Growth theorists make.

Kaganovich shows that constant returns in production need only hold asymptotically, that is, in the long run. Asymptotically the dynamics of the Kaganovich model are the same as the dynamics of the Dasgupta and Mitra model. The growth rate of consumption and capital stocks is determined in just the same way as in the simple Ak model of Sergio Rebelo (1991) with the von Neumann growth rate replacing the coefficient A . The condition that characterizes optimal balanced growth may be found by differentiating the simple Euler equations, based on the reduced utility function in discrete time. No fancy mathematics from the calculus of variations is needed, much less an appeal to Pontryagin. To determine the growth rate of an optimal path, first, take the product of the von Neumann maximum rate of growth times the discount factor. Then raise the product to a power equal to the elasticity of intertemporal substitution. Asymptotically this is the growth rate of the optimal path. In addition, relative prices and interest rates are asymptotically independent of preferences, thus of discount rates on future utility. In view of these facts it is clear that, asymptotically, the composition of consumption and the ratios of inputs into productive processes are constant. These results are reminiscent of the arguments of Sraffa (1960) as one might expect when human capital is treated as an output, since these models really do represent production of commodities by means of commodities.

Other convergence theorems have been proved in the New Growth literature in special models of endogenous growth. These use either one- or two-sector models. Some of the models have had competitive equilibria that are Pareto optimal. Others have been competitive but feature monopolistic competition or external economies where the competitive equilibrium is not Pareto optimal. In the case of competitive equilibrium that is Pareto Optimal, the convergence is to a balanced path of optimal capital accumulation which is optimal,

given the welfare function determined by the market equilibrium. However, the proofs of these results have involved features special to the dimensions of the models. Moreover, the linearization of the equations governing the evolution of the optimal paths have mostly led to matrices that are singular. For example, one of the capital goods has usually been produced without involving the other capital good, directly or indirectly. That is, human capital or education is produced without the introduction of physical capital. On the other hand, if the matrices that give the local approximation for the Euler equations for discrete time are assumed to be nonsingular, the balanced-growth paths of the endogenous-growth models with any number of capital goods are easily proved to be locally stable when the discount factors are near enough to 1.

So far as I know, interesting conditions that guarantee global stability for the case of any number of capital goods have not been found for the case of joint production. Of course, the no-joint-production model seems to include all the models that have been described in the New Growth Theory literature, in which the labor supply is not treated as a primary resource. If leisure is introduced into the utility function, I would say this condition is violated. So it is not surprising that multiple balanced paths arise, some of which are unstable. There is an excellent account of this literature in the January 1997 issue of the *Journal of Economic Dynamics and Control*. To have a proof of the turnpike in a truly general convex model it is necessary to eliminate the no-joint-production condition. Then the theorem would match Radner's proof of a Samuelson turnpike for the von Neumann model. Such a theorem seems to me to be within reach.

X. The Question of Foresight

A weakness of the turnpike theory in the setting of competitive equilibrium is the requirement of long-term foresight by the economic agents. It was first shown by Edmond Malinvaud (1953) that efficiency in production over time is not guaranteed by period-by-period maximization of profits, since capital overaccumulation could occur. More recently Frank Hahn (1966) in particular has empha-

sized that satisfying the conditions of market equilibrium in the short run is no guarantee that the economy is on an optimal path. It is well known that the prices that guide the choices of the economic agents must be such that a transversality condition at infinity is satisfied. However, it was suggested by Gale in his classic paper (1967 p. 2) that continual revisions of the plan might be able to overcome the deficiencies of long-range foresight: "to describe the situation figuratively, one is guiding a ship on a long journey by keeping it lined up with a point on the horizon even though one knows that long before that point is reached the weather will change (but in an unpredictable way) and it will be necessary to pick a new course with a new reference point, again on the horizon rather than just a short distance ahead."

Attempts have been made to prove theorems where short-term foresight is sufficient. In the context of planning for optimal growth this program was begun by Steven Goldman (1968) who showed that planning for a finite program which was revised every period would converge to the optimal program in a one-sector Ramsey style neoclassical model if the planning period was chosen to be sufficiently long. He required that the terminal capital stock at the end of the planning period be at least as large as the capital stock at the time of planning. This result has been generalized to multisector models since then.

These attempts to reduce the demands on foresight do not deal with the questions of technical progress and the introduction of new goods. A recent unpublished and, indeed, unfinished paper by Boldrin and David Levine entitled "Innovation Growth and Cycles in General Equilibrium" approaches these problems by way of a model with an infinite number of goods and activities. Only a finite set of goods and activities are present at any one time. Their utility functions are defined in terms of characteristics, rather than goods. They remark, "... convergence to any balanced growth path is at best partial and temporary, as new feasible activities are implemented when they become profitable and others are disbanded when not profitable anymore." Their arguments are interesting and provocative, but it is not possible for me to discuss here.

In the end, the growth models, like all economic models, are guides to the kinds of things that may happen. They cannot predict with the accuracy expected of the natural sciences. Of course the refined accuracy of prediction of the natural sciences is realized principally in controlled situations in the laboratory or in industry, with perhaps the singular exception of the movements of the celestial bodies. In the context of demographic theory Nathan Keyfitz (1977 p. 86) remarks, "An exposition of the mathematics of population is not more directly concerned with prediction of future changes than a book on hydrodynamics is concerned with the prediction of floods. The most one can hope is that theoretical formulations will give the practitioners who do the predicting some help in thinking about their problem." I suppose we should not expect the position of the theoretical economist to be stronger than that of the theoretical demographer.

REFERENCES

- Atsumi, Hiroshi. "Neoclassical Growth and the Efficient Program of Capital Accumulation." *Review of Economic Studies*, April 1965, 32(2), pp. 127-36.
- . "The Efficient Capital Programme for a Maintainable Utility Level." *Review of Economic Studies*, July 1969, 36(3), pp. 263-87.
- Barro, Robert and Sala-i-Martin, Xavier. *Economic Growth*. New York: McGraw-Hill, 1995.
- Becker, Gary S.; Murphy, Kevin M. and Tamura, Robert. "Human Capital, Fertility, and Economic Growth." *Journal of Political Economy*, October 1990, 98(5), part II, pp. S12-S37.
- Becker, Robert A. "The Equivalence of a Fisher Competitive Equilibrium and a Perfect Foresight Competitive Equilibrium in a Multi-sectoral Model of Capital Accumulation." *International Economic Review*, February 1982, 23(1), pp. 19-34.
- Becker, Robert A. and Boyd, John H., III. *Capital theory, equilibrium analysis and recursive utility*. Malden, MA: Blackwell, 1997.
- Bewley, Truman. "An Integration of Equilibrium Theory and Turnpike Theory." *Journal of Mathematical Economics*, September 1982, 10(2/3), pp. 514-40.
- Boldrin, Michele and Woodford, Michael. "Equilibrium in Models Displaying Fluctuations and Chaos: A Survey." *Journal of Monetary Economics*, March 1990, 25(2), pp. 189-222.
- Brock, William A. and Gale, David. "Optimal Growth under Factor Augmenting Progress." *Journal of Economic Theory*, October 1969, 1(3), pp. 229-43.
- Brock, William A. and Mirman, Leonard J. "Optimal Economic Growth and Uncertainty: The Discounted Case." *Journal of Economic Theory*, June 1972, 4(3), pp. 479-513.
- Cass, David. "Optimum Growth in an Aggregative Model of Capital Accumulation: A Turnpike Theorem." *Econometrica*, October 1966, 34(4), pp. 833-50.
- Cass, David and Shell, Carl. "The Structure and Stability of Competitive Dynamical Systems." *Review of Economic Theory*, February 1976, 12(1), pp. 1-10.
- Coale, Ansley J. "The Effects of Changes in Mortality and Fertility on Age Composition." *Milbank Memorial Fund Quarterly*, July 1957, 34(3), pp. 79-114.
- Dolmas, Jim. "Endogenous Growth in Multi-sector Ramsey Models." *International Economic Review*, May 1996, 37(2), pp. 403-21.
- Dorfman, Robert; Samuelson, Paul A. and Solow, Robert. *Linear programming and economic analysis*. New York: McGraw-Hill, 1958.
- Friedman, Milton. *A theory of the consumption function*. Princeton, NJ: Princeton University Press, 1957.
- Gale, David. "On Optimal Development in a Multi-sector Economy." *Review of Economic Studies*, January 1967, 34(1), pp. 1-18.
- Goldman, S. M. "Optimal Growth and Continual Planning Revision." *Review of Economic Studies*, April 1968, 35(102), pp. 145-54.
- Hahn, Frank. "Equilibrium Dynamics with Heterogeneous Capital Goods." *Quarterly Journal of Economics*, November 1966, 80(4), pp. 633-46.
- Kaganovich, Michael. "Sustained Endogenous Growth with Decreasing Returns and

- Heterogeneous Capital." *Journal of Economic Dynamics and Control*, 1998 (forthcoming).
- Keyfitz, Nathan. *Introduction to the mathematics of population with revisions*. Reading, MA: Addison-Wesley, 1977.
- Koopmans, Tjalling C. "On the Concept of Optimal Economic Growth," in *The Econometric Approach to Development Planning*, Pontificae Academiae Scientiarum Scripta Varia No. 28. Amsterdam: North-Holland, 1965, pp. 225–87.
- Lopez, Alvaro. *Problems in stable population theory*. Princeton, NJ: Office of Population Research, 1961.
- Lucas, Robert E., Jr. "On the Mechanics of Economic Development." *Journal of Monetary Economics*, July 1988, 22(1), pp. 3–42.
- Lucas, Robert E., Jr. and Stokey, Nancy L. "Optimal Growth with Many Consumers." *Journal of Economic Theory*, February 1984, 32(1), pp. 139–71.
- Malinvaud, Edmond. "Capital Accumulation and Efficient Allocation of Resources." *Econometrica*, April 1953, 21(2), pp. 233–68.
- McKenzie, Lionel W. "The Turnpike Theorem of Morishima." *Review of Economic Studies*, October 1963a, 30(2), pp. 169–76.
- . "Turnpike Theorems for a Generalized Leontief Model." *Econometrica*, January–April 1963b, 31(1–2), pp. 165–80.
- . "Accumulation Programs of Maximum Utility and the von Neumann Facet," in J. N. Wolfe, ed., *Value, capital, and growth*. Edinburgh: Edinburgh University Press, 1968, pp. 353–83.
- . "Turnpike Theorems with Technology and Welfare Function Variable," in J. Los and M. Los, eds., *Mathematical models in economics*. New York, Elsevier, 1974, pp. 271–87.
- . "Turnpike Theory." *Econometrica*, September 1976, 44(5), pp. 841–65.
- . "Turnpike Theory, Discounted Utility, and the von Neumann Facet." *Journal of Economic Theory*, August 1983, 30(2), pp. 330–52.
- . "Optimal Economic Growth, Turnpike Theorems and Comparative Dynamics," in K. J. Arrow and Michael D. Intriligator, eds., *Handbook of mathematical economics*, Vol. III. Amsterdam: North-Holland, 1986, pp. 1281–1355.
- Mirrlees, James A. "Optimum Growth When Technology Is Changing." *Review of Economic Studies*, January 1967, 34(1), pp. 95–124.
- Morishima, Michio. "Proof of a Turnpike Theorem: The No Joint Production Case." *Review of Economic Studies*, February 1961, 28(1), pp. 89–97.
- Radner, Roy. "Paths of Economic Growth That Are Optimal with Regard Only to Final States." *Review of Economic Studies*, February 1961, 28(1), pp. 98–104.
- Ramsey, Frank P. "A Mathematical Theory of Saving." *Economic Journal*, December 1928, 38(152), pp. 543–59.
- Rebelo, Sergio. "Long-Run Policy Analysis and Long-Run Growth." *Journal of Political Economy*, 1991, 99(3), pp. 500–21.
- Romer, Paul M. "Increasing Returns and Long-Run Growth." *Journal of Political Economy*, October 1986, 94(5), pp. 1002–37.
- . "Endogenous Technological Change." *Journal of Political Economy*, October 1990, 98(5), Part 2, pp. S71–S102.
- Samuelson, Paul A. *Market mechanisms and maximization*, Part III. Santa Monica, CA: Rand, 1949.
- Scheinkman, José A. "On Optimal Steady States of n -sector Growth Models when Utility Is Discounted." *Journal of Economic Theory*, February 1976, 12(1), pp. 11–30.
- Solow, Robert M. "A Contribution to the Theory of Economic Growth." *Quarterly Journal of Economics*, February 1956, 70(1), pp. 65–94.
- . *Growth theory, an exposition*. Oxford: Clarendon, 1970.
- Solow, Robert and Samuelson, Paul A. "Balanced Growth under Constant Returns to Scale." *Econometrica*, July 1953, 21(3), pp. 412–24.
- Sraffa, Piero. *Production of commodities by means of commodities: Prelude to a critique of economic theory*. Cambridge: Cambridge University Press, 1960.

- Swan, Trevor W. "Economic Growth and Capital Accumulation." *Economic Record*, November 1956, 32(44), pp. 334–61.
- Takahashi, Harutaka. "Characterization of Optimal Programs in Infinite Horizon Economies." Ph.D. dissertation, University of Rochester, 1985.
- Uzawa, Hirofumi. "Optimum Technical Change in an Aggregative Model of Economic Growth." *International Economic Review*, January 1965, 6(1), pp. 18–31.
- . "Time Preference and the Penrose Effect in a Two-Class Model of Economic Growth." *Journal of Political Economy*, July/August 1969, 77(4), Part 2, pp. 628–52.
- von Neumann, John. "Über ein Ökonomisches Gleichungssystem und eine Verallgemeinerung des Brouwerschen Fixpunktsatzes." *Ergebnisse eines Mathematischen Kolloquium*, 1937, 8, pp. 73–83; translated in "A Model of General Economic Equilibrium." *Review of Economic Studies*, 1945, 32, pp. 85–104.
- Yano, Makoto. "The Turnpike of Dynamic General Equilibrium Paths and Its Insensitivity to Initial Conditions." *Journal of Mathematical Economics*, December 1984, 13(3), pp. 235–54.
- . "On the Dual Stability of a von Neumann Facet and the Inefficacy of Temporary Fiscal Policy." *Econometrica*, March 1998, 66(2), pp. 427–52.

CLIO AND THE ECONOMIC ORGANIZATION OF SCIENCE[†]

Common Agency Contracting and the Emergence of "Open Science" Institutions

By PAUL A. DAVID*

The Cold War's ending has brought mounting pressures to recognize national science and technology research systems. Yet, by comparison with what has been learned already concerning institutional arrangements and business strategies affecting corporate R&D investments, surprisingly little is known about the economic origins and effects of the corresponding institutional infrastructures shaping the world of "academic" science, and the organization and conduct of publicly supported R&D more generally. The desirability of closing this particular lacuna in the economics and economic-history literatures has been just as evident to economists concerned with extending the analysis of modern institutions as to those who have begun to approach the whole area of science and technology studies from the perspectives and methods of industrial-organization economics.¹ Even before the "new economics of science" had

begun to direct attention to such a program, Douglass North (1990 p. 75) saw a significant challenge and a promising opportunity in explicit exploration of "the connecting links between institutional structures ... and incentives to acquire pure knowledge." The research reported here has accepted that challenge (see also the other papers in this session: Timothy Lenoir [1998], Christophe Lécuyer [1998], and Marjory S. Blumenthal [1998]). It is focused upon key episodes in the institutional evolution of "public science," and its complex and changing relationship to the other organizational spheres of contemporaneous scientific activity: those in which research was conducted under "proprietary rules" for industrial profit-goals, and "defense-related" science and engineering knowledge was sought under conditions of restricted access to information concerning methods, findings, and their actual and potential applications.

I. The Problem: Why "Open Science"?

The particular historical development of interest here is the emergence of precisely those fundamental lines of cultural and institutional demarcation, to which I have referred in distinguishing the existence of the sphere of "open science" activities supported by state funding and the patronage of private foundations and carried on today in universities and public (not-for-profit) institutes. Although the conceptualization of science as the pursuit of "public knowledge" now seems to many a natural, even a primitive notion, it is in reality a complex social construct (see Robert K. Merton 1973, 1996 part III). The "communal" ethos and norms of "the Republic of Science" emphasize the cooperative character of the larger purpose in which individual researchers are engaged, stressing that the

[†] Discussants: Kenneth Flamm, Brookings Institution; Zvi Griliches, Harvard University; David Mowery, University of California-Berkeley.

* All Souls College, Oxford OX1 4AL, U.K., and Department of Economics, Stanford University, Stanford, CA 94305-6072. This condensed presentation draws on David (1998), which should be consulted for refinements, qualifications, historical documentation, and references to the relevant literature. I am grateful to Avner Greif, Mario Biagioli, Partha Dasgupta, Weston Headley, Scott Mandelbrote, Joel Mokyr, Noel Swerdlow, and many other colleagues, institutions, and foundations whose intellectual and material generosity aided my research in this area since 1991. The comments and suggestions of the discussants improved the present version, although they did not make the task of compression any the easier.

¹ With particular reference to "the new economics of science," see Partha Dasgupta and David (1987, 1994), David (1994a), and the more recent surveys by A. M. Diamond (1996), Paula Stephan (1996), and David et al. (1998).

accumulation of reliable knowledge is an essentially social process. The force of its universalist norm is to render entry into scientific work and discourse open to all persons of "competence," while a second key aspect of "openness" is promoted by norms concerning the sharing of knowledge in regard to new findings and the methods whereby they were obtained.

Open science is a quite recent social innovation, at least by historical standards. Accompanying the profound epistemological reorientation wrought by the fusion of experimentalism with Renaissance mathematics, the cultural ethos and social organization of Western European scientific activities during the late 16th and 17th centuries underwent a significant transformation, a break from the previously dominant regime of "secrecy in the pursuit of nature's secrets." This change should be seen as a distinctive and vital aspect of the Scientific Revolution, from which there crystallized a new set of conventions, incentive structures, and institutional mechanisms that reinforced scientific researchers' commitments to rapid disclosure and wider dissemination of their discoveries and inventions. Yet the puzzle of why and how this came about has not received the notice it would seem to deserve, especially in view of the complementarities and tensions that are present today in relations between the regimes of open and proprietary science.

Any familiarity with the antecedent intellectual orientation and social organization of scientific research in the West would be sufficient to suggest the utter improbability of that historical bifurcation, which saw a new and quite antithetical mode of conducting the search for knowledge emerge alongside (and in some sense in competition with) the older, secretive hunting of nature's secrets. Medieval experimental science was shaped by a political and religious outlook that encouraged withholding from the "vulgar multitude" arcane and occult knowledge that might impart immense powers over material things (see William Eamon, 1994). The imperative of secrecy was particularly strong in the medieval and Renaissance traditions of alchemy, and it persisted there side-by-side with the emergent institutions of open science throughout the 17th and

into the 18th century. Social and economic regulations during the Middle Ages, along with rent-seeking strategies, worked in the same direction: knowledge of recently discovered geographical secrets and maps indicating trade routes would be closely guarded. Similarly, technological recipes were kept from the public domain by craftsmen, even when they were not compelled by guild restrictions to preserve the "mysteries" of the industrial arts.²

Why then, out of such a background should there have emerged a quite distinctive community of inquiry whose members came to be governed by a distinctive reward system based upon priority (and hence, necessarily, the revelation) of discoveries? Why, especially when in the modern context we see few if any differences between the methods of (scientific) inquiry used by university researchers working under the institutional norms of open science and the procedures that they (or others with the same training) employ in the setting of a corporate R&D laboratory? Can the social organization of open science then be simply an epiphenomenon of the philosophical and religious changes that some cultural historians see as underpinning the Scientific Revolution, if not of the epistemological transformations that it constituted? Stating the problem more synthetically, is it not plausible that these two discontinuities, the one taking place in the social organization of scientific inquiry and the other transforming its intellectual organization, were interdependent and entangled with each other in ways that need to be more thoroughly understood?

Considering the economic logic of the organization of knowledge-producing activities provides a start toward answering this question; it is possible to give a complete functionalist account of the institutional complex that

² From the 14th century to the early 18th century in Europe, the issuance of "letters patent" and granting of royal "privileges" conferring monopoly rights in exchange for the disclosure of technological information were aimed primarily to effect the transfer and application of existing industrial arts and engineering practices (i.e., techniques already known to master-craftsmen and engineers in other territories), and not particularly at inducing fresh inventive activity (see David, 1993a).

characterizes modern science in such terms (see Dasgupta and David, 1987, 1994). In brief, the norm of openness is incentive-compatible with a collegiate reputational reward system based upon accepted claims to priority; it also is conducive to individual strategy choices whose collective outcome reduces excess duplication of research efforts and enlarges the domain of informational complementarities. This brings socially beneficial spillovers among research programs and abets rapid replication and swift validation of novel discoveries. The advantages of treating new findings as public goods in order to promote the faster growth of the stock of knowledge are thus contrasted with the requirement of restricting informational access in order to enlarge the flow of privately appropriable rents from knowledge stocks. This functionalist juxtaposition suggests a logical basis for the existence and perpetuation of institutional and cultural separations between two normatively differentiated communities of research practice: the open "Republic of Science" and the proprietary "Realm of Technology" are distinctive organizational regimes, each of which serves a different (and potentially complementary) societal purpose.

The foregoing, "logical-origins" style of explanation for the institutions of modern science (and technology), however, is unconcerned with the details of their actual historical evolution. A rationale of this kind, at best, seems to presuppose a creationist fiction, namely, that these arrangements were instituted *ab initio* by some external agency, such as an informed and benevolent political authority endowed with fiscal powers. A response to that objection requires probing for the historical origins of the institutions of open science, since these remain outside the set of logical origins arrived at simply from a consideration of the present-day functional value of an extant, cooperative mode of scientific research.

II. The Argument: Noble Patrons, Mathematicians, and Principal-Agent Problems

I contend that the historical emergence of the norms of disclosure and demonstration and the rise of "cooperative rivalries" in the rev-

elation of new knowledge (the "open-science revolution") had independent and antecedent roots. These lay in the social and institutional contexts in which the new breed of scientists were working: the formation of a distinctive research culture of open science was first made possible and, indeed, was positively encouraged by the system of aristocratic patronage prevailing in an era when kings and nobles (both lay and ecclesiastical) were immediately concerned with the ornamental benefits to be derived by their sponsorship of philosophers and savants of great renown. To sustain this interpretation I argue that the economic logic of the patronage system in post-Renaissance Europe induced the emergence and promoted the institutionalization of new reputation-building proceedings; these entailed the revelation of scientific knowledge and expertise among extended reference groups that included "peer-experts." Patronage, however, already was an old system in the 17th century, for the sponsorship of intellectuals was a long-standing prerogative and responsibility of Europe's social and political elites. It is necessary then for me to explain that something new had appeared on the scene at that particular juncture; something which by disturbing that system induced the primitive formation of conventions and norms that can be identified with open science. The core part of my proposed explanation derives from considering, first, the economics of patronage in general, and then the specific implications of the newly arising problems of principal-agent contracting that were created by the late-Renaissance patronage system's encounter with the new (mathematical) form of "mechanical philosophy," in which the likes of Galileo, Johannes Kepler, and their contemporaries came to be engaged.

Aristocratic patronage systems historically reflected two kinds of motivation: the *utilitarian* and the *ornamental*. Most political elites, in addition to recognizing some need in their domain for men capable of producing new ideas and inventions to solve mundane but nonetheless important problems, also have sought to enlist the services of those who profess an ability to reveal the secrets of Nature, and of Destiny. Kings, princes, and lesser nobles sought to surround themselves with

creative talents whose achievements would enhance their self-esteem and their public image. Thus, poets, artists, musicians, chroniclers, architects, instrument-makers, and natural philosophers found employment and protection in aristocratic courts, both because their skills might serve the necessities and pleasures of the court and because their presence "made a statement" in the quest among nobles for prestige. Patron-client relations often were precarious, being uncomfortably subject to the volatility of aristocratic tastes and moods, and to the abrupt terminations that might ensue on a patron's disgrace or demise. Nonetheless, these dyadic connections existed in this era as part of a well-articulated system characterized by elaborate conventions and rituals that provided calculable career paths for men of intellectual and artistic talents (see Bruce Moran, 1991; Mario Biagioli, 1993).

Those motives for extending patronage as symbolic acts of public self-aggrandizement are here subsumed under the heading "ornamental." Such reasons should be understood to have been no less instrumental in their nature and roots than were the utilitarian grounds for the patronage of intellectuals. Grandeur and ostentatious display could serve to reinforce the claim of a prince to rightful possession of authority; the public display of "magnificence," in which art and power were closely allied, was a stock item in the repertoire of Renaissance state-craft (see Roy Strong, 1984). This is significant, because inventions and discoveries that met utilitarian needs in many instances would have to be kept secret if they were to be most useful, whereas it is in the nature of the ornamental motive for the patronage of creative talent that its fulfillment elicits the disclosure of new, marvelous discoveries and productions—that the client's achievement on behalf of the patron be widely publicized. Indeed, it was very much in the interest of a patron for the client's reputation to be enhanced in this way, for the fame of the latter augmented his own.

Into this setting a new element had been interjected during the 16th century. The more extensive and rigorous use of mathematical methods formed an increasingly important aspect of the work of natural philosophers and

others.³ A side-effect of this intellectual advance was, however, to render the basis of the mathematically sophisticated savants' claims and reputations less immediately accessible for evaluation by the elites in whose service they wished to be employed. The difficulties thus posed by the asymmetric distribution of information were rather unprecedented, not having been encountered to the same degree in the patronage of intellectuals and artists who followed other, less esoteric callings. The new breed of scientists, however, claimed to specialize in revealing the unfamiliar. Opportunities for charlatany here were more rife, and so were the risks of embarrassment for the patron, should it turn out that one had sponsored a fraud—or much worse, a heretic. Thus, even where the services of the mathematically trained *intelligencia* might be sought for essentially practical, utilitarian pursuits (designing machinery for public spectacles, surveying and cartography, ballistics, or the correct use of perspective in pictorial arts), the soundness of the candidates' qualifications had become more problematic and far from inconsequential.

This shift was tantamount to the emergence of especially compelling reasons for noble patrons readily to delegate more of the responsibility for evaluating and selecting among the new breed of savants; those screening functions were thereby devolved initially to informal networks of correspondents, and increasingly to more institutionalized communities of their fellow practitioners and correspondents. Except for those few who were themselves adepts, patrons were inclined to refrain from passing personal judgment on scientific assertions, or involving themselves in substantive controversies (see Biagioli, 1993). It was left to the initiative of the parties dependent upon such patronage to organize the production of credible testimonials to their own credibility and scientific status. Not alto-

³ See C. B. Boyer (1985 Ch. 15) and Noel Swerdlow (1993) on Renaissance mathematics; A. Keller (1985) on the program and rhetoric developed on behalf of mathematical training during the 1570's and 1580's; and M. Feingold (1984) and Biagioli (1989, 1993) on the patronage of mathematicians.

gether surprisingly then, the beginning of the era of modern mathematics also witnessed the formation of active networks of correspondence among Europe's algebrists and geometers, announcing newly devised techniques and results; the mid-16th century initiated the tradition of publicly posing mathematical puzzles, issuing scientific challenges, announcing prizes for the solutions of problems, and the holding of open competitions to test the claims of rival experts in the mathematical arts. On the interpretation proposed here, the new practices of disclosure constituted a functional response to heightened asymmetric-information problems that the mathematization of natural philosophy and the practical arts posed for the Renaissance system of court-patronage.

III. Common Agency Contracting, with Rival Principals—The Legacy of European Feudalism

The foregoing sketch of the early modern court patronage system presents features recognizable to economists as those of "common agency contracting," involving the competition of incompletely informed rival principals for the dedicated services of an expert agent. Establishing that correspondence suggests three significant propositions about the economic organization of scientific activities in Europe during the late 16th and early 17th centuries.

First, since what the scientist-clients had to offer their patrons was "novelty," at any point in time the welfare of a scientist's several patrons could not be jointly advanced by the same degree. In the early history of modern science, as a consequence of the dominance of patrons who were concerned with the ornamental rather than the utilitarian value of scientist-philosophers, the services a client provided to his several patrons were more in the nature of *positional goods*, and hence essentially were "substitutes" rather than "complements."

Second, in the majority of cases, the material rewards offered to clients by any single patron were not sufficiently large and certain to relieve them from the quest for multiple patrons. But in the absence of full information and concerted action on the part of principals, the nature of the incentive contracts offered by

the latter would have reflected their awareness of the possibility that a client/agent could use the means provided by one patron to serve the ends of another. Under these common agency conditions the resulting Nash equilibrium in the game among rival principals would be a set of patronage-contracts that offered clients comparatively weak material incentives to devote their efforts exclusively to the service of any one patron (see Avinash Dixit, 1996). Such an outcome, of course, would be consistent with the necessity of seeking to serve a number of patrons concurrently for "piece-meal compensation." Even though a scientist such as Galileo might deplore that situation as demeaning (Biagioli, 1993 p. 29), it worked nevertheless to reinforce would-be clients in their adoption of research and publication strategies that widened the circle of their reputation.

Third, as Lars Stole's (1990) analysis of mechanism design under common agency contracting has shown, the equilibrium outcome in the case of "contract substitutes" is in general more favorable to the agent than is the case when the services performed for different principals are complements. In effect, the competition among patrons to command the faithful attention of an agent/client (when they cannot free-ride on the knowledge products delivered to their rivals) leads to incentive structures that allow the client to retain more of the "rents" from the specialized information he possesses. The situation therefore tended to provide greater rewards for scientific activities than would have been the case otherwise, were there only a single possible patron on the scene; or had the patrons predominantly enjoyed positive externalities from others' support of the agent's efforts. The latter, of course, is the characteristic situation when there are significant *spillovers* of (utilitarian) benefits from new knowledge.

In the story related here there is an historical irony well worth remarking upon, especially as it underscores the tenacity of the past's hold on the incrementally evolving institutions that channel the course of economic change.⁴ Here

⁴ On "path dependence" in the dynamics of economic systems, see, for example, David (1993b, 1994b, 1997).

is the nub of it: an essentially precapitalist, European aristocratic disposition to engage in the patronage of intellectuals of renown for ornamental motives came to confer value upon those who pursued knowledge by following the new science in the late 16th and 17th centuries. The norms of cooperation and information disclosure within the community of scientists, and their institutionalization through the activities of formal scientific organizations, emerged (in part at least) as a response to the informational requirements of a system of patronage in which the competition among noble patrons for prestigious clients was crucial. Those rivalries were a legacy of western European feudalism: the medieval fragmentation of political authority had set the stage for common agency contracting in *substitutes*. An instructive contrast might be drawn with the alternative circumstances of a monolithic political system, such as had prevailed elsewhere, as in the Heavenly Empire of China during an earlier epoch, to cite a well-known case of a society that clearly possessed the intellectual talents for great scientific accomplishments, yet failed spectacularly to institutionalize the practice of open science. Might one then see open science to have been European feudalism's great gift to the economic vigor of capitalism in the modern age?

IV. Conclusion

Some important part of the impact of science today derives from the radical social innovation that the open-science regime constituted. A corollary proposition, to which the historical experience recounted here also lends support, is that the methods of modern science in and of themselves were not and still are not sufficient to form the unique cultural ethos associated with the "Republic of Science." Nor can they be expected automatically to induce and sustain the peculiar institutional infrastructures and organizational conditions of the open-science regime, within which their application has proved so conducive to the rapid growth of the stock of reliable public knowledge and all that flows therefrom. Rather than being the robust epiphenomena of a new organum of intellectual inquiry, the institutions of open science are independent and in some

measure fortuitous social and political constructs; along with the cultural ethos they have served to transmit from generation to generation, they are in reality intricate legacies of European history.

Features of the institutional infrastructure of public science, being thus in some significant degree exogenous to actual scientific practice, may be subjected to substantial redesign and otherwise manipulated as potent instruments of state science and technology policies. In one sense that is the good news. But it comes with a caution. While the norms of openness play a critical part in maintaining the systemic efficacy of modern scientific research, they are terribly vulnerable to the withdrawal of public-minded patronage and protection. Wise policy-making in this critically sensitive area must pay especial heed to the complex and contingent histories of the organizations of public science and so respect the potential fragility of the peculiar institutional matrix within which modern research evolved and has flourished.

REFERENCES

- Biagioli, Mario. "The Social Status of Italian Mathematicians." *History of Science*, 1989, 27, pp. 41–95.
- . *Galileo, courtier: The practice of science in the culture of absolutism*. Chicago: University of Chicago Press, 1993.
- Blumenthal, Marjorie S. "Federal Government Initiatives and the Foundations of the Information Technology Revolution: Lessons from History." *American Economic Review*, May 1998 (*Papers and Proceedings*), 88(2), pp. 34–39.
- Boyer, C. B. *A history of mathematics*. Princeton, NJ: Princeton University Press, 1985.
- Dasgupta, Partha and David, Paul A. "Information Disclosure and the Economics of Science and Technology," in G. Feiwel, ed., *Arrow and the ascent of modern economic theory*. New York: New York University Press, 1987, pp. 519–42.
- . "Toward a New Economics of Science." *Research Policy*, 1994, 23, pp. 487–521.
- David, Paul A. "Intellectual Property Institutions and the Panda's Thumb: Patents,

- Copyrights, and Trade Secrets in Economic Theory and History," in M. Wallerstein, M. Magee, and R. Schoen, eds., *Global dimensions of intellectual property protection in science and technology*. Washington, DC: National Academy Press, 1993a, pp. 19–61.
- . "Path-Dependence and Predictability in Dynamic Systems with Local Network Externalities: A Paradigm for Historical Economics," in D. Foray and C. Freeman, eds., *Technology and the wealth of nations: The dynamics of constructed advantage*. London: Pinter, 1993b, pp. 209–31.
- . "Positive Feedbacks and Research Productivity in Science: Reopening Another Black Box," in O. Grandstrand, ed., *Technology and economics*. Amsterdam: Elsevier, 1994a, pp. 65–89.
- . "Why Are Institutions the 'Carriers of History'? Path Dependence and the Evolution of Conventions, Organizations and Institutions." *Structural Change and Economic Dynamics*, 1994b, 5(2), pp. 205–20.
- . "Path Dependence and the Quest for Historical Economics." University of Oxford Discussion Papers in Economic and Social History No. 20, November 1997.
- . "Reputation and Agency in the Historical Emergence of the Institutions of 'Open' Science." University of Oxford Discussion Papers in Economic and Social History No. 23, February 1998.
- David, Paul A.; Foray, Dominique and Steinmueller, W. Edward. "The Research Network and the New Economics of Science: From Metaphors to Organizational Behaviors," in A. Gambardella and F. Malerba, eds., *The organization of innovative activities in Europe*. Cambridge: Cambridge University Press, 1998 (forthcoming).
- Diamond, A. M., Jr. "The Economics of Science." *Knowledge and Policy*, Summer/Fall 1996, Special Issue, 9(2/3), pp. 6–49.
- Dixit, Avinash. *The making of economic policy: A transaction cost politics perspective*. Cambridge, MA: MIT Press, 1996.
- Eamon, William. *Science and the secrets of Nature: Books of secrets in medieval and early modern science*. Princeton, NJ: Princeton University Press, 1994.
- Feingold, M. *The mathematicians' apprenticeship: Science, universities and society in England, 1560–1640*. Cambridge: Cambridge University Press, 1984.
- Keller, A. "Mathematics, Mechanics, and the Origins of the Culture of Mechanical Invention." *Minerva*, 1985, 23(3), pp. 348–61.
- Lécuyer, Christophe. "Academic Science and Technology in the Service of Industry: MIT Creates a 'Permeable' Engineering School." *American Economic Review*, May 1998 (*Papers and Proceedings*), 88(2), pp. 28–33.
- Lenoir, Timothy. "Revolution from Above: The Role of the State in Creating the German Research System, 1810–1910." *American Economic Review*, May 1998 (*Papers and Proceedings*), 88(2), pp. 22–27.
- Merton, Robert K. *The sociology of science: Theoretical and empirical investigations* [N. W. Storer, ed.]. Chicago: University of Chicago Press, 1973.
- . *On social structure and science* [P. Sztompka, ed.]. Chicago: University of Chicago Press, 1996.
- Moran, Bruce, ed. *Patronage and institutions: Science, technology, and medicine at the European court, 1500–1750*. Woodbridge, Suffolk, U.K.: Boydell, 1991.
- North, Douglass C. *Institutions, institutional change and economic performance*. New York: Cambridge University Press, 1990.
- Stephan, Paula. "The Economics of Science." *Journal of Economic Literature*, 1996, 34(3), pp. 199–235.
- Stole, Lars. "Mechanism Design Under Common Agency." Working paper, Massachusetts Institute of Technology, 1990.
- Strong, Roy. *Art and power*. Woodbridge, Suffolk, U.K.: Boydell, 1984.
- Swordlow, Noel. "Science and Humanism in the Renaissance: Regiomontanus's Oration on the Dignity and Utility of the Mathematical Sciences," in P. Horwich, ed., *World changes: Thomas Kuhn and the nature of science*. Cambridge, MA: MIT Press, 1993, pp. 131–68.

P-11028

Revolution from Above: The Role of the State in Creating the German Research System, 1810–1910

By TIMOTHY LENOIR*

Discussions of modern scientific research's organization point to the 19th-century emergence of German research universities as evidence that state investment in nondirected academic research, when coupled with beneficial relations between academic research and industry, and when stimulated by appropriate incentives such as protection of intellectual property in an open, competitive system, can lead to explosive growth in scientific knowledge and rapid improvement of industry. This paper examines three episodes in the evolution of Germany's research system pointing to the roles of state interests and innovative ministerial leadership in fashioning the research system to meet state needs.

I. Economic Engineering: 1806–1848

The first period considered is that bracketed by the Prussian defeat by Napoleon's armies and the 1848 Revolution. Discussions of the research university quite naturally focus on the Humboldt reforms and the University of Berlin's founding in 1810. But the Humboldt reforms should be considered together with the Stein-Hardenberg economic reforms aimed at fostering private initiative through removing guild restrictions on trade as well as a sweeping set of anti-feudal land and labor reforms. In a nearly unbroken line of policy until 1845, Karl H. F. Stein, Karl A. F. Hardenberg, and their successors, Gottlieb J. C. Kunth, Ludwig F. V. Bülow, and especially Christian Beuth and Christian Rother, attempted to stimulate industrial development. They employed a variety of means, from dissemination of technical information to handing over government-purchased foreign technology to private parties as capital investment, and from

the creation of organizations to generate development funds for new industrial start-ups to actively building state-financed industries employing the newest technologies and organizational techniques, all in order to pressure Prussian industry to modernize its production methods (see William D. Henderson, 1958; Ulrich Peter Ritter, 1961; Ilja Mieck, 1965; Wolfgang Radtke, 1981; Hubert Kiesewetter, 1989; Hermann Fernholz, 1991; Werner Vogel, 1993).

Another key feature of the Prussian plan to modernize German industry was new types of educational institutions to free industry from its tradition-bound practices. Stein and Kunth identified "Bildung" as the most useful form of state aid, and Beuth argued that where science is not introduced into industry, there can be no securely based industry or progress. Beuth opened the Gewerbeschule (School of Trade and Industry) in Berlin in 1821, to provide rudimentary instruction to handworkers and manufacturers in mechanics and chemical-technical subjects. The school was expanded within a few years to include a third class, the "Suprema," which treated the scientific basis of technology as a unified field of study.

The Humboldt university reforms were also conceived as regenerating the nation's spiritual foundation, particularly through institution of the seminar, nourishing intellectual independence and initiative. Elite young minds trained in close interaction with faculty working on independent research problems would become the new generation's bulwark. But the university Humboldt envisioned did not include laboratory training as a regular part of the science curriculum. In fact, Humboldt's planned science curriculum was primarily devoted to "pure" theoretical science, particularly mathematics and physics. While chemistry and physiology were included in this picture, laboratories were only deemed

* Department of History, Stanford University, Stanford, CA 94305.

important for supporting lecture demonstrations and for the professor's private research needs. The model of *Wissenschaft um sich selber willen* that emerged in this environment was heavily opposed to any association with handwork. Between 1830 and 1848, laboratory exercises were tolerated as entrepreneurial activities professors might initiate for fees which they would plow back into lab equipment (R. Steven Turner, 1971, 1974; Charles McClelland, 1980).

The response of the Prussian ministry and university faculty to a scathing critique of the system launched in 1840 by Giessen chemist Justus Liebig sums up the pre-1848 situation of natural sciences in universities. To Liebig's claim that chemistry was an independent discipline worthy of its own institute, rather than simply an adjunct field for medical students, leading faculty members responded that Liebig's recommendations to combine the pursuit of new chemical knowledge with laboratory work based on standardized and easily taught methods of analysis undermined the university's purpose. Liebig's program combined pursuit of pure knowledge, typical of science academies, with work appropriate to technical institutes, which trained students in material production. According to an old-guard professor, Liebig personified the time's central academic evil: lust after discovery in order to attract more students. The true purpose of university science, according to this professor, was to transmit solid, proven knowledge to men training in useful professions to serve the state (Turner, 1982).

II. Research Imperative, Decentralized Competition, and Institute-Building: 1848–1871

A fundamental shift occurred in the organization of academic science in Germany from the mid-1840's through the mid-1870's, connected with the emergence of a prestige market driven by a new research ethos (Joseph Ben-David, 1971). This shift in many ways realized Liebig's vision of science. As Ben-David argued persuasively, in large part the shift during this period was due to competition among different German states for intellectual talent as they vied for cultural leadership of a hoped-for unified Germany. Intense competi-

tion existed among the leading state ministries of culture and education to stock their universities with the best professors, now defined as discoverers of new knowledge. A state ministry's appointments were based on international reputation for research and publication as evaluated by a review process established within the ministry and drawing on faculty peer review. In an environment where several universities could compete for a single professor's talents, highly visible scientists were able to make laboratory space, assistants, and equipment a condition of their acceptance. These academic market forces meant that nearly every German university got at least a small institute of chemistry, and similar developments occurred in physics and physiology. Having grown out of traditional teaching functions in which laboratory work was seen as at best formalizing traditional student training, once the new labs were in place entrepreneurial directors were recruited who would use the facilities to advance their own research and to encourage research among advanced students, in turn reinforcing the scientific achievement of the professor/lab director.

While the open system blossomed during this brief period of German academic history, the earlier mercantilistic concerns of state ministries were nonetheless still present. In my view, interaction between these two tendencies gave the system its distinctive features. Recruiting star faculty was only one of the items on state ministers' agendas during this period, characterized at best by modest economic growth and more often by stagnation and decline. Enhancing their universities' prestige was an important goal, but increasingly for the smaller German states, stimulating their economies was another. While universities competed for faculty, they also competed with one another for students; a reason for recruiting star faculty was to increase student enrollments. Since the largest growth sector of student matriculation in this middle period was in medicine, ministers tried to increase their competitive advantage for medical students by hiring the faculty and constructing the ancillary support facilities for medical education. A second, larger, area of concern of state ministers during the late 1840's–1850's was the need to stimulate the economy. While

states like Prussia could appoint a few professors in highly visible universities without consideration of the support facilities' integration with other programs, smaller states felt increasing pressure to utilize their resources efficiently, dovetailing appointments with other initiatives. Peter Borscheid's (1976) study of Baden illustrates how this competition for students, combined with the effort to harness chemistry to stimulate agricultural production, led to impressive expansion of chemistry facilities at the University of Heidelberg.

Elsewhere (Lenoir, 1997), I have shown that a similar pattern can be seen in the building of first-class science and medical institutes at the University of Leipzig by Johannes Falkenstein, the director of the Saxon Kultusministerium. Falkenstein was to Saxony what Beuth and Rother were to Prussia (see David Cahan, 1985; Alan J. Rocke, 1993; Lenoir, 1997). As Kultusminister of the most industrialized German state, Falkenstein was not as pressed as his Baden colleagues to generate immediate economic benefit to industry and agriculture from investment in the natural sciences, but he had consistently promoted economic modernization and industrialization as a means to long-term prosperity and social stability. But one of his concerns as Kultusminister was to boost enrollment at Leipzig, particularly by attracting students from other German and foreign states. Leipzig had not done particularly well in student enrollment during the 1840's and early 1850's. In an era when the medical faculty was a university's "bread and butter," Falkenstein concentrated his resources in the area of his existing strength: clinical medicine. He made plans to improve the Leipzig natural-science faculties and build a new science and medical campus.

The retirement of the professor of physiology allowed Falkenstein to implement his plan. He recruited Carl Ludwig, Germany's leading physiologist, for Leipzig's clinical medicine program. Although Ludwig was famous for introducing physics-based instrumentation into physiology, his work had never had much direct contact with clinical medicine. Falkenstein perceived a perfect fit between Ludwig's advancing research programs and the work of the star he already had

in the Leipzig medical stable, the clinician Karl Wunderlich. Since the late 1850's, Wunderlich had been deeply involved in his studies on thermometry, strengthening his conviction that the closest cooperation ought to develop among experimental physiology, chemistry, pathological anatomy, and diagnostic techniques for the clinic. Falkenstein built new institutes for physiology and pathology, each with positions for assistants in physiological chemistry and microscopic anatomy, the goal being to integrate these various enterprises into a collaboration between Ludwig and Wunderlich. In order to promote this cooperation, these institutes were all situated adjacent to one another with connecting corridors and walkways. The integration of experimental clinical medicine, long considered a defining moment of the 19th century's medical revolution and the first step toward rational science-based medicine, can be considered the outcome of strategic planning on the part of enlightened state bureaucrats (see Kiesewetter, 1988).

The efforts of ministers like Falkenstein to optimize interactions among their different research faculty and to coordinate facility use led to systemic interactions among disciplines, improving the content of science in ways that no one could have predicted by simply betting on each discipline's stars pursuing their own personal research programs. While this system building was considerably short of targeted research and development, it prepared the way.

III. Academic Science and Industry's Needs: 1877-1910

The final phase of development I want to consider is between 1871 and 1910, culminating in the formation of the Kaiser-Wilhelm Institutes. As the two periods I have already discussed demonstrate, the view that state-supported research at the universities should stimulate industry in certain ways was at best a rhetorical position in the German nation's political and cultural transformation. For the most part, few scientists other than Liebig thought their work had direct relevance to industry, and the highest rewards were to be obtained by emerging as "bearers of culture," rather than as scions of industry. In the period

between 1871 and 1910, however, this situation shifted radically, when the tensions that had earlier characterized the relations between academic and industrial cultures dissolved. This cultural shift was as important as the increased relevance of scientific research to the economic performance of German industry. In bringing about this transformation, the work of enlightened state ministers, particularly Friedrich Althoff, was once again crucial (see Karl-Heinz Manegold, 1970; Frank R. Pfetsch, 1974; Bernhard vom Brocke, 1980).

In the previous sections I have described the conditions for a powerful set of institutions for generating scientific research and the production of new knowledge. But by 1890 the recognition dawned that the fusion of teaching and research providing the rationale for developing these institutions in fact hindered science's advance, since the bulk of resources had to go into supporting time-consuming low-level training. The progress of science at the turn of the century depended on more than simply providing good scientists with time for research by reducing their teaching. In addition, I have suggested in the previous section that developments such as those connected with Ludwig's call to Leipzig pointed toward another refinement of the system through stimulating interconnections between otherwise autonomous disciplines. Although favorably disposed to such cross-fertilization, Althoff did not hold out much prospect for its realization, given the rigid hierarchies and social divisions in German universities. The solution to these structural problems depended on establishing an Archimedean point outside the universities. The impetus for change and the specific solutions to the problems of establishing the necessary institutional conditions for advanced research came about as a result of acute awareness of the increasing importance of academic science for industry, and the centrality of industry to the German Imperial State.

The relationship between Farbwerke Hoechst, a dyestuffs manufacturer, and the research workers in Robert Koch's laboratory at the Imperial Institutes of Health (*Reichsgesundheitsamt*), following the discovery in Koch's lab of the first biological antitoxins (diphtheria, tetanus) illustrates how closer

links between industry's demands and academic science's interests were forged. A constellation of factors stimulated research and development in the field of pharmaceuticals at Hoechst. Foremost among these was the economic crisis faced by the industry, most severe between 1881 and 1885. The increased costs of finding new dyes in an evermore competitive market and dyestuffs' actual decline in price forced major chemical firms producing dyes, such as Hoechst, to seek to diversify their products (John Joseph Beer, 1959; Borscheid, 1976).

Hoechst moved into pharmaceuticals by establishing consulting arrangements with the directors of university laboratories, supplying them with both funding and materials to work on subjects of potential interest to the firm. Within a few years, owing to the increasing complexity of research requiring facilities unavailable routinely in universities and problems of reliably managing such consulting arrangements, firms like Hoechst moved to internalize these research capabilities. Early experiences with the privatization of Emil Behring's work on sera at Hoechst led Althoff to be concerned about insuring the expansion of basic research, but also making it available for commercial development. A first experiment was the establishment of the Frankfurt Institut für Serum-Prüfung-und-Forschung with Paul Ehrlich as director. Though not yet a for-profit research institute, the Institut was firmly based on mutual cooperation among state, industry, and academic science. This institute prefigured a more ambitious endeavor to pursue rational drug therapeutics and production of artificially synthesized drugs based on Ehrlich's discoveries of certain dyes' ability to block trypanosomes' toxic capabilities (see Ernst Bauemler, 1984; Lenoir, 1997 pp. 179–202).

The Georg-Speyer-Haus that Ehrlich proposed and eventually constructed was an interdisciplinary institution whose director would define problems to be attacked through exchange of ideas among physiologists, biochemists, microbiologists, bacteriologists, pharmacologists, and clinicians working in-house. An unspecified percentage of the profits from patents was reinvested in the institute to cover its operating costs, including the costs

of undertaking new research. The firms of Hoechst and Casella contributed substantially to the initial endowment and also supplied the raw materials used in the department of chemistry's research. In exchange, the two firms received first refusal on any marketable patents. But the choice of research problems was left to be determined solely by Ehrlich and his staff.

Concrete results, such as the production of Salvarsan, the first cure for syphilis, were undoubtedly instrumental in supporting the conviction that the pattern of a multidisciplinary institute combining the advancement of basic science with the needs of industry embodied in the Georg-Speyer-Haus was capable of more general application. It was because of its success that Ehrlich sat on the board of advisers who laid the plans for the Kaiser-Wilhelm-Gesellschaftinstitutes (see Manfred Rasch, 1987; Jeffrey Allan Johnson, 1990; Rudolf Vierhaus and vom Brocke, 1990).

REFERENCES

- Bauemler, Ernst. *Paul-Ehrlich: Scientist for Life*. New York: Holmes and Meier, 1984.
- Ben-David, Joseph. *The scientist's role in society: A comparative study*. Englewood Cliffs, NJ: Prentice Hall, 1971.
- Beer, John Joseph. *The emergence of the German dye industry*. Urbana: University of Illinois Press, 1959.
- Borscheid, Peter. *Naturwissenschaft, Staat und Industrie in Baden, 1848–1914*. Stuttgart: Klett, 1976.
- Cahan, David. "The Institutional Revolution in German Physics, 1865–1914." *Historical Studies in the Physical Sciences*, 1985, 5(2), pp. 1–65.
- Fernholz, Hermann. *Beuth: Deutsches Bürger-tum vor 100 Jahren: Dem Verein zur Förderung des Gewerbflusses, 1821–1931*. Berlin: Verein zur Förderung des Gewerbflusses, 1991.
- Henderson, William O. *The State and the Industrial Revolution in Prussia, 1740–1870*. Liverpool, U.K.: Liverpool University Press, 1958.
- Johnson, Jeffrey Allan. *The Kaiser's chemists: Science and modernization in Imperial Ger-many*. Chapel Hill: University of North Carolina Press, 1990.
- Kiesewetter, Hubert. *Industrialisierung und Landwirtschaft: Sachsens Stellung im regionalen Industrialisierungsprozess Deutschlands im 19 Jahrhundert*. Vienna: Boehlau, 1988.
- . *Industrielle Revolution in Deutschland, 1815–1914*. Frankfurt: Suhrkamp, 1989.
- Lenoir, Timothy. *Instituting Science: The Cultural Production of Scientific Disciplines*. Stanford, CA: Stanford University Press, 1997.
- Manegold, Karl-Heinz. *Unversitaet, Technische Hochschule und Industrie: Ein Beitrag zur Emanzipation der Technik im 19 Jahrhundert unter besonderer Beruecksichtigung der Bestrebungen Felixs Kleins*. Berlin: Duncker and Humblot, 1970.
- McClelland, Charles. *State, society, and university in Germany, 1700–1914*. New York: Cambridge University Press, 1980.
- Mieck, Ilja. *Preussische Gewerbepolitik in Berlin, 1806–1844*. Berlin: DeGruyter, 1965.
- Pfetsch, Frank R. *Zur Entwicklung der Wissenschaftspolitik in Deutschland, 1750–1914*. Berlin: Duncker and Humblot, 1974.
- Radtke, Wolfgang. *Die Preussische Seehandlung zwischen Staat und Wirtschaft in der Frühphase der Industrialisierung*. Berlin: Colloquium Verlag, 1981.
- Rasch, Manfred. *Vorgeschichte und Gruendung des Kaiser-Wilhelm-Instituts fuer Kohlenforschung in Mulheim a.d. Ruhr*. Hagen, Germany: Linnepe, 1987.
- Ritter, Ulrich Peter. *Die Rolle des Staates in denfruehstadien der Industrialisierung*. Berlin: Duncker and Humblot, 1961.
- Rocke, Alan J. *The quiet revolution: Hermann Kolbe and the science of organic chemistry*. Berkeley: University of California Press, 1993.
- Turner, R. Steven. "The Growth of Professorial Research in Prussia, 1818–1848—Causes and Context." *Historical Studies in the Physical Sciences*, 1971, 3(2), pp. 137–82.
- . "University Reformers and Professorial Scholarship in Germany, 1760–1806," in Lawrence Stone, ed., *The university in society*, Vol. 2. Princeton, NJ: Princeton University Press, 1974, pp. 495–531.

- _____. "Liebig and Prussian Chemistry: Reflections on Early Institute-Building in Germany." *Historical Studies in the Physical Sciences*, 1982, 13(1), pp. 129–62.
- Vierhaus, Rudolf and vom Brocke, Bernhard. *Forschung im Spannungsfeld von Politik und Gesellschaft: Geschichte und Struktur der Kaiser-Wilhelm-Max-Planck-Gesellschaft: Aus Anlass ihres 75 jährigen Bestehens*. Stuttgart; Deutsche Verlags-Anstalt, 1990.
- Vogel, Werner. *Die Seehandlung: Preussische Staatsbank: Handel, Verkehr, Industrie, Bankwesen*. Berlin: Geheimes Staatsarchiv Preussischer Kulturbesitz und der Stiftung Preussische Seehandlung, 1993.
- vom Brocke, Bernhard. "Hochschul und Wissenschaftspolitik in Preussen und im Deutschen Kaiserreich, 1882–1907: Das 'System Althoff,'" in Peter Baumgart, ed., *Bildungspolitik in Preussen zur Zeit des Kaiserreichs*. Stuttgart: Klett-Cotta, 1980, pp. 9–118.

Academic Science and Technology in the Service of Industry: MIT Creates a “Permeable” Engineering School

By CHRISTOPHE LÉCUYER *

The Massachusetts Institute of Technology had an exceptional trajectory from the 1880's to the early 1920's. From a struggling “Polytechnic Institute” in the early 1880's, it became within less than 40 years the only independent engineering school in America that performed significant research and reached national stature. Unlike most institutions of higher education, MIT became, also, a “permeable” institution closely linked to industry—a school where industrial contacts were institutionalized and expanded to a degree unequaled by other universities and where servicing industry became very much a part, and even a central part, of the institutional culture.¹

Focusing on the conflicts among different coalitions of faculty members and administrators who had divergent and sometimes opposed views about MIT's future, its educational policies, and its proper relations with industrial pursuits, this essay analyzes the contested transformation of the Institute into a “permeable” institution closely allied to industrial corporations during the progressive era. More broadly, this paper examines the relations between academic institutions and industrial corporations from the early 1880's to the mid-1920's.

MIT's relationship to industry, often considered representative and paradigmatic of academia's partnership with business, has been studied by David Noble (1977), John Servos (1980), and W. Bernard Carlson (1988). Noble has argued that during the progressive era exigencies of corporate capitalism had a profound influence on the engineering profession and led to attempts by

corporate engineers to “design a new social order, one dominated by the private corporation and grounded upon the regulated progress of science and technology” (Noble, 1977 p. xxiv). Without significant opposition, corporate engineers built this new socio-technical order by appropriating scientific technology and by transforming schools such as MIT into “units of the industrial system” (Noble, 1977 p. 169). This provocative thesis has been criticized by Servos and Carlson. These authors have contended that the integration of science into American corporations was neither effortless nor orderly. More particularly, they have questioned Noble's claim that academic institutions became handmaidens to business interests in the 1910's and 1920's. In particular, focusing on a bitter conflict in MIT's Chemistry and Chemical Engineering Department between faculty members in favor of a close partnership with industry and those who seemingly opposed it, Servos (1980) argued that corporate interests dominated the Institute only momentarily.

New material on the 1880's and 1890's, a period overlooked by Noble, Servos, and Carlson, leads to a rethinking of the university not as the product of large scale forces or as the outcome of a single vision, but as the result of the competing programs of different faculty groups. Unlike previous studies, this paper argues that MIT's curricula, research programs, and institutional culture as well as its close relations with industry were the materialization of heterogeneous and often conflicting programs. It also claims that corporate patrons played a secondary—if significant—role in the evolution of the university. By supporting certain faculty groups instead of others, industrial patrons altered the balance of power at the Institute. But this does not mean that they single-handedly fashioned MIT's institutional culture, curricula, and research programs. Furthermore, I argue that MIT's case does not

* History Department, Building 200, Stanford University, Stanford, CA 94305-2024.

¹ For a longer version of this paper, see Lécuyer (1995).

support the larger claim of Noble, Servos, and Carlson that science and technology were generally integrated into corporate America during the progressive era. Instead, I contend that a more adequate picture should emphasize a limited process of harmonization between academic and industrial practices and the establishment of fragile alliances between groups within academia and between academia and industry.

I. Conflicting Projects

A struggling polytechnic school during the 1860's and 1870's, MIT grew into one of the best-known and largest undergraduate engineering schools in America during the last two decades of the 19th century. Central to MIT's rise to prominence was a group of practical engineers and Yankee reformers who developed a "democratic" and "practical" form of engineering education. Critical of classical studies which "served the persistent conservatism of a privileged class" and of schools such as Harvard that offered them, they wanted to bring engineering to the people and built an engineering school which served students of modest means (Jacob Bigelow, 1865). They also built a practically oriented curriculum that fitted students for immediate usefulness in industry. As a way of keeping abreast with professional practice and finding jobs for the graduates, they forged close ties with local industrial corporations. Regular visits to factories were organized to acquaint students and faculty with current engineering practice. Faculty members also engaged in consulting. Charles Cross, for instance, the head of the physics and electrical engineering programs, was the chief expert of the Bell Telephone Company in the 1880's and early 1890's. Finally, many departments appointed practicing engineers as lecturers to teach the latest technical developments.

During the second half of the 1890's and increasingly during the 1900's, MIT administrators and faculty members confronted a series of problems including heavy financial deficits, rising competition from Harvard and Midwestern universities, and the emergence in professional societies of a new concept of the engineer as an expert and manager which

threatened MIT's preeminent place in engineering education. As these problems festered, the practical engineers' leadership was increasingly questioned by presidents Henry Pritchett (1900-1907) and Rupert Maclaurin (1909-1920) as well as by three new faculty groups that often participated in different progressive movements: the science faction around physical chemist Arthur Noyes, the engineering-reformers around William Walker and Dugald Jackson, and a coalition of second-generation practical engineers led by chemist Henry Talbot. Intent upon meeting rising competition from research universities and Midwestern engineering schools, they sought to expand the Institute, introduce a stronger element of science into the curricula, and develop research and graduate programs. They also favored closer links with industry. But because these administrators and faculty members held different views of engineering and had often distinct political agendas, they gave different meanings to "industrial service" and advocated divergent projects for the Institute.

Whereas Pritchett, a Roosevelt progressive and a promoter of America's industrial leadership, sought to centralize MIT and reform it thoroughly through a merger with Harvard, Maclaurin, in accord with his ideology of industrial supremacy, wanted to build a "school of science that concerned itself with practical affairs" (Maclaurin, 1916). The science faction around Noyes, who thought of engineering as laboratory science, advocated science-centered engineering curricula, promoted the establishment of research laboratories along the German model, and wanted to transform MIT into a science-based research university. They also sought to forge intimate links with industry. In particular Noyes promoted collaborations between the faculty and the research laboratories of large science-based corporations such as General Electric and DuPont.

Competing with the science faction, the engineer-reformers who shared Maclaurin's industrial nationalism, wanted to reform engineering education in accord with their view of engineering as management and transform MIT into an elite technological school closely connected to industrial pursuits. In particular, Walker, a chemical engineer, advocated the

establishment of an MIT-centered network of industrial-research laboratories which would serve small companies and make American industry more competitive and scientific. Only through "a closer alliance between the scientific worker and the actual agencies of production," Walker reasoned, could the country "find its industrial salvation," improve the "often mean and sordid" living conditions of the working classes, and provide a "general moral and spiritual uplift" to the "great masses of the community" (Walker, 1911).

Positioned in the middle of the practical engineers, the science faction, and the engineer-reformers, a coalition of pragmatically oriented faculty members around chemist Talbot sought to "modernize" the Institute without breaking with its tradition of preparing students for positions of "immediate usefulness" in industry. As a result, this new breed of practical engineers embraced elements of the reform programs of Walker and Noyes, such as the establishment of research laboratories and the institutionalization of relations with industrial corporations. But whereas, for Walker, research laboratories were a tool for transforming small industrial corporations and making American industry more competitive, for Talbot they were mainly a more efficient way of achieving what the first generation of practical engineers had done in the 1880's and 1890's: keeping in touch with engineering practice and finding jobs for the graduates.

II. The Making of a "Permeable" Engineering School

Repeated struggles and shifting alliances among Pritchett, Maclaurin, the practical engineers, and the groups led by Noyes, Walker, and Talbot shaped the Institute and transformed it into a permeable engineering school closely linked to industry. Even though these groups all left their mark on the school, some were much more influential than others. The practical engineers, who had set MIT's trajectory and controlled most departments, largely blocked the reformers' initiatives until the mid-1910's. Pritchett, who disregarded MIT's institutional culture, lost nearly every battle

and saw his merger project with Harvard lead to an all-out faculty revolt. Noyes and his group played a significant role in the institutionalization of research. In particular, in 1903 they established the Research Laboratory of Physical Chemistry, which trained industrial researchers and gained a prominent place in physico-chemical research in America. Although Noyes and his group controlled the presidency between 1907 and 1909, they failed to transform MIT into a research university, for of all the reforming programs, theirs was the most alien to MIT's "practical" and "democratic" tradition. Only alliances with other coalitions, the enlisting of patronage, and the lasting control of departmental chairmanships would have helped the science faction overcome the resistance of many members of the faculty.

Conversely, it was largely because they held key positions in the academic structure, forged external alliances, and benefited from the retirement of many practical engineers that Maclaurin, the Walker faction, and the group led by Talbot were able to expand MIT and partially reorient its course during the 1910's and 1920's. In alliance with the engineer-reformers and the second generation of practical engineers, President Maclaurin embarked between 1909 and 1920 upon a program of institutional advancement which gave a central place to industrial "service." To construct a school that would contribute to the making of an "orderly" and "scientific" industrial society, Maclaurin enlisted the patronage of George Eastman of Eastman Kodak. Becoming MIT's largest patron, Eastman gave the Institute more than \$10 million between 1912 and 1919.

Eastman's lavish patronage reversed MIT's financial fortunes. It also made possible the construction of a new campus in Cambridge and the school's rapid expansion. Eastman's support also helped implement many of the reforms advocated by Walker's group and by the second generation of practical engineers. It financed the development of new undergraduate curricula along the lines advocated by the Walker and Talbot groups as well as the building of graduate programs in engineering. As a result, conferral of graduate degrees grew more than eightfold, from 21 in 1910 to 168

in 1924. By the mid-1920's, MIT awarded one-third of the country's master's degrees in engineering and more than one-half of its engineering doctorates.

In conjunction with their educational reforms, the Maclaurin, Walker, and Talbot groups institutionalized relations with industry and transformed MIT into an "instrument of research regarding the problems of industry" (Maclaurin, 1917 p. 21). The engineer-reformers and the second generation of practical engineers created new institutional mechanisms such as industrial advisory committees, cooperative teaching programs, and industry-supported laboratories in electrical, chemical, and metallurgical engineering. Walker's Chemical Engineering Practice School, founded with Eastman's support, is a case in point. A network of industrial research laboratories, the Chemical Engineering Practice School was designed to make industry "more scientific" and to "serve the small company which could not afford a staff of research men" (Warren K. Lewis, 1927). In residence at each of the stations was a member of MIT's teaching staff who instructed the students and at the same time conducted research of interest to the host company.

Enlarging their campaign for "industrial research" to the Institute as a whole, Maclaurin and the engineer-reformers established the Technology Plan in 1919. Centrally administered by the Division of Industrial Cooperation and Research, the Technology Plan offered industrial corporations access to libraries, alumni records, and faculty consulting against an annual retaining fee. The Technology Plan raised considerable interest in industry, attracting no fewer than 200 corporate subscribers. The contributors included science-based corporations such as GE and AT&T and large steel and rubber companies. But there were also numerous smaller concerns in the textile, machinery, and paper-making industries.

As a result of these and other efforts, MIT's industrial contacts were institutionalized and expanded to a degree unequaled by other universities by the mid-1920's. More importantly, "industrial service" became a central part of the MIT's institutional culture. Sponsored research, still exceptional during the 1910's, became common practice after the war: research

support increased by 500 percent from \$56,452 in 1920 to \$264,797 in 1927.² By that date, it was estimated that more than one-third of the staff was actively engaged in research, testing, and commercial analyses for industry. Similarly, faculty consulting grew considerably. Most engineering department chairs directed "downtown" consulting firms. More than half of the staff regularly consulted for outside concerns during the 1920's. More importantly, the unstated rules that governed the institution changed. After the war, faculty appointments and nominations to department chairmanships increasingly depended upon consulting and development work for industrial concerns.

III. Conclusion

MIT's history between 1890 and 1920 cannot be reduced to the conflict between Noyes and Walker studied by Servos; nor can it be construed as the product of corporate capitalism or the best example of corporate control over academic institutions. A closer analysis of the relationships among academic administrators, faculty members, and industrial patrons reveals a far more complex and interesting story. MIT's anomalous trajectory can be largely explained by the institution's unusual makeup and by the conflicts and shifting alliances among various administrators and faculty groups, each having its own educational and institutional project that was often informed by specific political views and a particular ideology of engineering. Repeated struggles and shifting alliances among Pritchett, Maclaurin, and the faculty groups shaped the Institute and transformed it into a permeable engineering school in close contact with industry.

Although the Institute's interaction with industry was neither representative nor paradigmatic, a number of lessons can be learned from MIT's case. First, unlike what has been previously assumed, close university-industry relations were not new in the mid-1900's. At MIT, the coalition of practical engineers had encouraged them since the early 1880's. By

² For a study of sponsored research at MIT during the 1920's, see Leroy Foster (1984).

1900, a significant fraction of the staff consulted for industrial concerns. It might be noted, however, that a shift appears to have occurred between 1905 and 1910. Because the new faculty groups wanted to construct an efficient and orderly industrial society, they gave a much greater value to relations with industry. They institutionalized cooperation with industrial firms through novel organizational forms such as advisory committees and engineering research laboratories and, for the first time, attempted to enlist corporate patronage.

Second, this essay also questions the contention that corporate engineers took control of MIT and more generally "retooled American higher education 'processes' to meet industrial needs" (Noble, 1977 p. 130). MIT administrators and faculty groups shaped the school's policies without much industrial interference. They were also firmly in control of the curriculum and most research programs. But this does not mean that corporate engineers and industrial firms had no influence. To advance engineering research, administrators and faculty groups compromised and often assigned patent rights and final say over publication to corporate patrons. While Eastman does not seem to have had a powerful voice in the school's policies, his support of Maclaurin's and Walker's reforms deeply altered the balance of power at MIT and helped them implement their institutional and disciplinary programs.

Third, MIT's case does not seem to support larger claims that during the progressive era science and technology were integrated into business and that "engineering education" became "a unit of the industrial system" (Noble, 1977 p. 169). Instead of "integration" and the high degree of unification that it implies, what seems to have been at work was a complex and bounded process of adjustment between academic and corporate practices and the construction of sometimes fragile alliances between various groups at MIT and in industry. At least three types of university-industry collaboration can be distinguished. The coalition of practical engineers allied themselves with local corporations to find jobs for the graduates and produce technologists who could be of immediate usefulness to industry. They hired practicing engineers as lecturers and encouraged consulting with local firms as

a way of fostering the teaching competence of the staff. The science faction was connected to science-based industry and particularly research laboratories at GE, Cooper Hewitt, and DuPont. They trained industrial researchers and worked on questions "the industries were approaching but had not yet reached" (Dugald Jackson [no date]). Whereas the science faction's alliance with industry was informal and it positioned its research programs in advance of industrial laboratories, the Walker group institutionalized its cooperation with small corporations through research laboratories and the Chemical Engineering Practice School, sought corporate patronage, worked on pressing commercial problems, and interlaced its educational activities with corporate practices.

In short, MIT's transformation into a "permeable" educational institution was shaped mostly by struggles among different faculty groups and by various political agendas. From 1870 to 1920, the Institute was a site of intersection of numerous political movements seeking to reshape American society and industry through science, technology, and the reform of educational institutions. Such developments should remind economists and historians that engineering education and university-industry relations had political meanings and were fashioned by political projects. They also indicate the need for studies that examine, in a close-grained fashion, the political dimensions of technological education within a system of corporate capitalism where there was often disagreement about the ideal form of the polity and the precise relationship of engineering education to the economy and to corporations.

REFERENCES

- Bigelow, Jacob. *Address on the limits of education read before the Massachusetts Institute of Technology*. Boston, MA: E. P. Dutton, 1865.
- Carlson, W. Bernard. "Academic Entrepreneurship and Engineering Education: Dugald C. Jackson and the MIT-GE Cooperative Engineering Course, 1907-1932." *Technology and Culture*, July 1988, 29(3), pp. 536-67.
- Foster, Leroy. "Sponsored Research at MIT, 1900-1968, Vol. 1." Unpublished manu-

- script, MIT Archives and Special Collections, 1984.
- Jackson, Dugald.** "Advanced Instruction and Research in the Electrical Engineering Sciences." MIT Archives and Special Collections MC5-4-270, no date (probably 1911 or 1912).
- Lécuyer, Christophe.** "MIT, Progressive Reform, and 'Industrial Service,' 1890–1920." *Historical Studies in the Physical and Biological Sciences*, 1995, 26(1), part 1, pp. 35–88.
- Lewis, Warren K.** "Conference between Heads of Departments and the Corporation on the Division of Industrial Cooperation and Research." Unpublished report, MIT Archives and Special Collections AC13-25-176L, 23 February 1927.
- Maclaurin, Rupert.** "A National Opportunity and a National Duty." *Stone and Webster Journal*, August 1916, 19, pp. 89–93; reprinted, MIT Archives and Special Collections AC13-13-385.
- . *Annual report of the President for the academic year 1915–1916*. Cambridge, MA: Massachusetts Institute of Technology, 1917.
- Noble, David.** *America by design*. New York: Oxford University Press, 1977.
- Servos, John.** "The Industrial Relations of Science: Chemical Engineering at MIT 1900–1939." *Isis*, September 1980, 71(3), pp. 531–49.
- Walker, William.** "Chemical Engineering and Industrial Progress." *Journal of Industrial and Engineering Chemistry*, May 1911, 3(5), pp. 286–92.

Federal Government Initiatives and the Foundations of the Information Technology Revolution: Lessons from History

By MARJORY S. BLUMENTHAL *

There is little argument that early investment in computing and communications (C&C) research and development (R&D) has paid off, given the growth and vitality of the computing industry in the United States. Less clear are how and why federal C&C research support has been effective overall; guidance for future research policy can come from an examination of history. Whereas a vigorous industry has been visible for decades, many outside observers who question the need for federal C&C research support are often ignorant of the continuing exploitation by industry of science and technologies developed decades ago via past support and how such support has complemented industrial activities. This paper outlines how the organization of federal funding for C&C R&D has helped to make such funding so fertile, noting dimensions that may be of enduring value even as circumstances change. It focuses on the High Performance Computing and Communications Initiative of the 1980's and early 1990's, which evolved from earlier funding programs and influences current programs and prospects.¹ By design (and reflecting space limitations), the paper emphasizes positive aspects that may be relevant for the future.

I. Whirlwind: Early Initiative with Impact

Computing as we recognize it today originated around World War II; Kenneth Flamm (1988) has documented the dominant drivers of military (and to a lesser extent other gov-

ernment) data processing needs beginning in the 1930's and 1940's. Military needs inspired funding to develop systems for such applications as cryptanalysis and fire control. Subsequently and through the Cold War, federal funding for relevant research evolved from an emphasis on direct funding of industry efforts to an emphasis on funding of university-based research. This broad trend reflects the rise of both a computer industry and an academic base for computer-science research, which were fostered by federal support for computing R&D.

Project Whirlwind illustrates the productive interplay of multiple organizations and sectors associated with early federal support for C&C R&D (Bruce Old, 1981; Flamm, 1988). Whirlwind emerged from early 1940's roots as a postwar project at the Massachusetts Institute of Technology, supported by the Navy (the Office of Naval Research [ONR], beginning in 1946). As its funding needs grew, so did the purported rationale (from flight simulation to a broader set of command and control applications). Controversies over its progress and growing funding requirements contributed to a shift to Air Force funding, with initial emphasis on air-traffic-control application. Resort to an alternative funding agency within the federal government is a factor that recurs in C&C history. The availability of multiple sources of funding allows continuity of effort and introduces new perspectives, both important to the eventual development of more and more general-purpose computing systems.

Whirlwind is reknowned in computing history because of its important contributions to three sectors: government, industry, and academia. It provided a technical basis for the semi-automated ground environment (SAGE) air defense system that was vital to technical progress at IBM, which worked with MIT and became the prime contractor for SAGE. Although it did not yield desired air defense ca-

* Computer Science and Telecommunications Board, National Research Council, 2101 Constitution Avenue, N.W., Washington, DC 20418.

¹ This paper draws on Computer Science and Telecommunications Board (1994) and a successor project, "Innovations in Computing and Communications: Lessons from History," for which a report will be completed in 1998. CSTB 1994 contains numerous references.

pabilities, it directly or indirectly contributed to the development of a number of key technologies (e.g., graphics, light pens, and displays; ferrite core memories; and programming and compiling) that became important in military and commercial computing. Whirlwind proved concepts that helped the industry to grow; it provided experience with technology and research management that enabled subsequent innovation and product development at IBM. It stabilized and enabled growth of an academic research team at MIT, it contributed to development of other nonindustrial research teams responsive to federal needs (the Lincoln Laboratory associated with MIT and the MITRE Corporation) that interacted with MIT researchers as well as industrial teams, and it provided the basis from which associated researchers went on to launch new industrial efforts (such as the Digital Equipment Corporation). It also catalyzed the development of computing-related teaching capabilities at MIT: IBM funding helped to equip and staff early (mid-1950's) teaching facilities at MIT, complementing ONR support for research, *per se*.

In short, through project Whirlwind, multiple funding and talent sources, flows of these resources across organizations and sectors, and adaptability of federal funding objectives contributed to desired and unanticipated innovation. In addition to development of specific capabilities useful to the government, academic capacity was launched that would benefit government and industry, and industry capacity was developed that would benefit government (and to some extent academia). The growth of industry capacity that accelerated in the 1960's provided supply and demand for research support that complemented and evolved with federal research support in the 1970's and later.

II. HPCCI: Recent Initiative with Impact

The High Performance Computing and Communications Initiative (HPCCI) illustrates the challenges of designing effective research support in the context of a broad and established but growing technology base, with commercial uncertainties shaped by perceived foreign competition and structural change in

the telecommunications industry. Absent the immediate government/military needs that impelled Whirlwind, HPCCI provided an integrated response to a variety of perceived military and civilian needs for advanced computing and communications. It blended three previously separate streams of federal support: R&D aimed at supercomputers, or more generally, high-performance computing, including hardware and software; R&D aimed at computer networking; and R&D aimed at computational science, or the joint advance of computing and some other science (e.g., physics) in which computing was becoming an important tool.²

HPCCI provided objectives and organizational structure within and among federal agencies beginning in 1989, although it became an official initiative in 1991. It arose from informal interactions among program managers at the Defense Advanced Research Projects Agency (DARPA), the National Science Foundation (NSF), the Department Of Energy (DOE), and the National Aeronautics and Space Administration (NASA), who had to convince science-policy leadership to support it.³ Their success was evident in a budget that grew to exceed \$1 billion and expansion to involve at least 12 agencies.

HPCCI built on program-manager interactions with members of the scientific research community in universities,⁴ many of whom expressed frustrations about the nature and quality of their computing and communications infrastructure given perceived potential; members of the computer-science research

² The discussion of HPCCI derives from Computer Science and Telecommunications Board (1994) as well as the author's experience with aspects of the initiative.

³ During the late 1980's the larger science policy-coordination structure was evolving; the Bush administration revived the Federal Coordinating Council on Science, Engineering, and Technology (FCCSET, now defunct) and promoted cross-agency science-based initiatives (e.g., in earth science/global warming, manufacturing). This period emphasized what has become a continuing trend of federal research encouragement of cross-disciplinary research.

⁴ By this time, several agencies (e.g., the Department of Energy, the National Aeronautics and Space Administration) ran computing centers used by academic researchers supported by those agencies.

community, who offered ideas for how computing and communications technologies could advance; and eventually from industry, which recognized that it benefited from academic research (as a source both of technology and of customers) and, sensitive to perceived foreign competitive challenges, sought to promote innovation most likely to support competitiveness.⁵ Both computational scientists and networking researchers (a subset of computer scientists) tend to claim paternity for HPCCI; each has deep roots in federal support and recognizes the demand for more and more capability by academic researchers. Among computer scientists, HPCCI is considered the first major federal initiative to emphasize the science of computing and communication, a sign that the field had come of age.

HPCCI promoted general and specific objectives for advancing U.S. leadership in computing and communications. The support for parallel processing enabled academic researchers to prove high-performance computing architecture concepts, and that paradigm has become commercialized successfully. The objective of achieving 1,000 times faster computing by the mid-1990's than that available at the beginning of HPCCI was as ambitious as objectives set by the federal government for computing power to meet military needs during and after World War II. The objective of network speeds of at least 1 gigabit per second (broadly deployed and accessible) also stretched capabilities. Complementing technological objectives was a goal of increasing the number of computer science Ph.D.'s.

Its multiple roots are responsible for the union within HPCCI of computer-science research and deployment of computer-science research results in computing and communications infrastructure. C&C research and deployment interact: software and application developers need to work on the most advanced systems possible to test new concepts, and hardware/architecture developers need to understand the needs and problems

facing software and application developers to aim their own contributions. A set of vehicles, within and across HPCCI agencies, fostered collaboration between computing and computational scientists. Emblematic of these vehicles were the Grand Challenge teams, typically both interdisciplinary and multi-institutional, which were funded to support research on such problems as weather forecasting and pollution modeling.

Over its lifetime, the HPCCI evolved along two, interrelated dimensions. First, the nature and mix of its substantive emphases changed. This was most notable in the expansion from a set of science-based "Grand Challenges" to a broader, additional set of "National Challenges"; this set was formalized with the High Performance Computing Act of 1991 (PL 102-194), which also reinforced the communications aspects of HPCCI. This development responded to a perceived need to make HPCCI seem more relevant to more people, given difficulties in justifying research funding levels to the Congress. It also broadened the cross-disciplinary collaborations fostered by HPCCI, linking computer scientists and experts in a variety of domains that might nominally have little emphasis on science research (e.g., libraries, education). Partnerships between academia and industry were an HPCCI subtheme, evident in support for specific developers of high-performance computing architectures, encouragement of industry cost-sharing for academic super-computer centers, and the development of teams behind several high-performance networking testbeds.

Second, a broader set of players emerged: among federal agencies, among university centers, and among the companies with relevant product interests. Some of this broadening occurred because of recognition, by those not originally involved, of the large amount of resources associated with HPCCI, some because of the more diverse set of substantive emphases. Third, the nature of the National Challenges implied a growing concern with problems of scale as well as speed in computing and communications—National Challenge arenas (e.g., manufacturing, health care) were by definition quite large.

⁵ During the late 1980's, trade associations were formed (e.g., the Computer Systems Policy Project) or reconceived to strengthen input to technology policy.

The concern for scale complemented growth in interest in networking. These interests were captured in the National Information Infrastructure (NII) initiative launched in 1993 (Information Infrastructure Task Force, 1993). Administration interest in the NII led to a new HPCCI component, the Information Infrastructure Technology and Applications component, which began officially in 1994.

The scale and number of participants in HPCCI made its leadership a source of contention. Some of the concern related to the amount of leadership: the first head of the National Coordinating Office, which was launched in 1992, served that role half-time, and the federal advisory committee authorized by the High Performance Computing Act of 1991, a potential source of perspective from industry and academia, was not established until the initiative's sunset. Some of the concern related to the kind of leadership: federal program managers, including NCO as well as component agency staff, staunchly argued for "coordination," although some outside observers from industry, in communicating with Congress and federal research agencies, asked regularly for coordination with more managerial clout.

HPCCI expired officially, after a five-year authorization period, in 1996. Federal program managers supported—in some cases, promoted—the ending of the initiative because changes in the Congress made the aggregation and integration that once had seemed virtuous now appear hazardous. Prior to the official expiration, the then-current cohort of federal program managers from the high-performance computing and communications community worked to transform the initiative from the inside out, positioning the allocation of research support in computer science (and to some extent in computational science) to build on the HPCCI, to emphasize elements relating to networking and information infrastructure, and to make support appear less monolithic. NCO continues, providing a meeting ground and various vehicles for program managers in multiple agencies to communicate and share information about similar, complementary, and potential programs. Experi-

mentation continues in fostering R&D partnerships among industry and academia via focused initiatives (e.g., Next Generation Internet) and the contraction from a handful of regional NSF-supported super-computer centers to two centers with nationwide networks of affiliates.

III. Constants and Changes: Lessons from Effective Initiatives

Contemporary research support faces many hurdles at a time when the very premises of federal support are under question (Linda Cohen and Roger Noll, 1994). Controversies over HPCCI and more recently its descendent Next Generation Internet (NGI) illustrate the difficulty of formulating federal research support programs in the context of steady commercialization of new technology. Like Whirlwind, HPCCI shows the value of pursuing ambitious and difficult technological objectives.

Successful federal research support has had a number of important characteristics: it has coupled potential consumers of advances with their developers, it has drawn on real problems and application contexts, it has extended over periods long enough to accommodate the challenges of building and testing real systems. The results have gone beyond the immediate products of specific projects to include unanticipated results deriving from use (e.g., electronic mail was not anticipated as a driving use of computer networking), it has fostered growth of the talent base and movement of talent among enterprises and sectors (and thereby effective knowledge propagation), and more. These traits are evident from examination of such early large-scale federal projects as Whirlwind and SAGE, and they recur in subsequent large-scale initiatives, including HPCCI. For example, the DARPA-NSF program to support very large-scale integrated circuit (VLSI) design introduced computer science into microelectronics to achieve significant gains in VLSI design. The program developed a new academic base of research and teaching, nurturing expertise in individuals who propagated knowledge by moving from such initial

bases as Stanford University and the California Institute of Technology to other universities and into industry, often forming new start-ups.⁶

Examination of effective federal research support through major initiatives illuminates key attributes. One is variety in sources of funds and type of program management. Because the C&C science and technology base is so diverse, and because research is needed across a broad range from theory to systems of varying scales, different kinds of project and program are needed. Agencies differ as to the kind of work they support effectively. For example, the support for research infrastructure (e.g., human resources, facilities) provided by NSF complements the support for networking protocol development provided by DARPA, and the support for network deployment and access to different groups of researchers by these and other agencies extended early network experimentation to a broader base than any one agency might have been able to support. Each agency has been managed with different styles and emphasis on different kinds of projects. NSF historically emphasized individual-investigator, small-scale research; DARPA has consistently emphasized multiple-researcher, large-scale research. Because of the difficulty of predicting what kind of C&C technology is really needed, let alone what will work, diversity in research support appears to have evolutionary benefit.

From the postwar period through the 1980's, C&C research support has benefited from visionary program managers. The instrumental contributions of the Office of Naval Research over decades can be attributed to the vision and judgment of Marvin Denicoff; the many fundamental advances associated with DARPA are linked to the vision and judgment of such scientists as J. C. R. Licklider, Ivan Sutherland, and Robert Kahn; more recently, at NSF Steven Wolff drove the networking research and infrastructure program that catalyzed the commercialization of the Internet. Such individuals had good ideas about where computer science and technology could go,

they could recognize good ideas when raised by researchers, and they worked by nurturing research communities explicitly, developing academic bases and encouraging university-industry interaction. Their contributions were recognized contemporaneously as well as in retrospect.

Beginning in the late 1980's, researchers have raised concerns about the quality and nature of research management. Growth in the industry and academic bases for computing make it harder for a program manager to know readily the best available talent and make it harder for agencies to attract program managers to federal jobs. Agency growth and steadily tightening oversight, meanwhile, seems to promote more bureaucratic program management. These conditions will complicate research support in the future.

In conclusion, unlike federal funding for classical physical sciences, federal funding for computer-science research served to launch and sustain a field over several decades. The success of that launch reflects multifaceted and adaptive organization of federal research support programs and dynamic interaction of the government with industry and academia. The influence of federal entities as consumers of desired technology has been strong and enduring, although it is diminishing and becoming more specialized as the government, like the rest of the economy, comes to depend more on commercial off-the-shelf technology. Speculative funding has also been a theme, gravitating over time to academic research support as industry has become more self-sufficient. Notwithstanding such good beginnings, the future of computer science is cloudy: industrial strength continues to build on innovations of varying vintages while intellectual leaders in the field are calling for new paradigms to deal with the problems emerging from past success and new conditions. Research successes have resulted in a vast space in which new innovation is possible, but there is much less vision than in earlier decades about where best and how to aim new efforts. The flexibility to recognize, respond to, and stimulate new ideas may be at least as important in the future as history has shown it has been in the past.

⁶ Personal communication via briefing, John Hennessy (Stanford University) and Charles Seitz (Myricom, Inc.), 5-7 February 1997.

REFERENCES

- Cohen, Linda R. and Noll, Roger. "Privatizing Public Research." *Scientific American*, September 1994, 271(3), pp. 72-77.
- Computer Science and Telecommunications Board, National Research Council. *Evolving the high performance computing and communications initiative to support the nation's information infrastructure*. Washington, DC: National Academy Press, 1994.
- Flamm, Kenneth. *Creating the computer: Government, industry, and high technology*. Washington, DC: Brookings Institution, 1988.
- Information Infrastructure Task Force. *The National Information Infrastructure agenda for action*. Washington, DC: Information Infrastructure Task Force, 15 September 1993.
- Old, Bruce S. (Bruce S. Old Associates, Inc.). *Return on investment in basic research—Exploring a methodology*. Office of Naval Research, Department of the Navy (Washington, DC) Contract N00014-79-C-0192, November 1981.

HISTORICAL PERSPECTIVES ON CURRENT ISSUES OF ECONOMIC PERFORMANCE[†]

Micro Rules and Macro Outcomes: The Impact of Micro Structure on the Efficiency of Security Exchanges, London, New York, and Paris, 1800–1914

By LANCE DAVIS AND LARRY NEAL*

When a formal securities market is established there are choices about the structure of the operating rules that must be made before the market can begin to function. Those choices may be made by the “owners” of the market or they may be made by governments. In either case, these rules constitute the micro structure of the securities market; and that structure, in turn, will have a substantial impact on the efficiency of the exchange in terms of costs, scope, volume, and level of penetration. This paper examines the differential impact of the rules governing the initial definition, and the enforcement, of property rights in the London, New York, and Paris securities exchange markets over the course of the long 19th century, 1800–1914.

In 1914, listings on the London Stock Exchange encompassed nearly one-third of the paid-up value of all negotiable securities in the world (£10.7 billion out of a total of £32.6 billion), and in every industrial category (government, railroad, or industrial and commercial) foreign issues exceeded domestic in both number and value. The New York Stock Exchange remained largely focused on American issues, but the listings were broad and the volume substantial. In 1913 the value of listed securities totaled some £5.4 billion, almost one-half (by value) of the securities in circu-

lation in the United States (U.S. Department of Commerce, 1975 p. 1007; Randal C. Michie, 1987 pp. 168, 195 [table 7.1]). In 1907 it is estimated that the total volume of securities quoted in Paris was £6.2 billion, an amount slightly larger than the figure for the New York exchange; but the Paris total includes the listings of both the Bourse and *Cou-lisse* (the “unofficial” exchange). Of that total, 58 percent of the listings were foreign (W. C. Van Antwerp, 1913 pp. 406–7). Evidence of the importance of initial property rights can be seen most strikingly in a comparison of the market prices of the right of access to each of the markets at the outbreak of World War I. The number of traders on the London Stock Exchange had risen from 1,076 in 1852 to 4,855 in 1914. In New York, the number had been kept by mutual agreement at 1,100 from 1879 to 1914. In Paris, the number of seats on the Bourse had been set by the government at 60 in 1801, a figure that was increased to 70 in 1898 (Van Antwerp, 1913 pp. 392–93; Michie, 1988 p. 60). Reflecting these differences in numbers, in 1914 the cost of a seat on the Paris Bourse was £92,000; in New York it was £16,000; and in London £1,200 for any “honourable man” or a mere £440 for anyone who had served as a broker’s clerk for four years. These differences were largely due to the set of property rights that were established early in each exchange’s operation.

I. How They Started

All three exchanges were established within ten years of each other (1792–1801) and for the same objective, improving the operation of

[†] *Discussants:* Michael Edelstein, Queens College, City University of New York; Richard Sylla, New York University.

* California Institute of Technology, Division of the Humanities and Social Sciences, 228-77, Pasadena, CA 91125, and University of Illinois, 1407 W. Gregory, 328A DKH, Urbana, IL 61801, respectively.

the secondary market for newly created national debt. Almost as if someone intended to create a controlled experiment on the effects of differing property rights, the assignment of rights in the new markets was quite distinct. In London, a private corporation with a limited number of shareholders (260) constructed a new building to house trading activity, mainly in government debt. By British common law, they were unable to exclude nonshareholders from the marketplace, so they deliberately set out to encompass all the possible traders and trade within the one exchange. They largely succeeded; there were 550 subscribers to the new facility (E. Victor Morgan and W. A. Thomas, 1962 p. 143). In New York, a much smaller number of brokers (24) agreed to trade only with each other. They also agreed to maintain minimum commissions charged to their clients. In Paris, Napoleon began to bring order out of the revolutionary chaos that had created open access to the stock exchange by restoring the government-enforced monopoly on trading in the reconstituted public debt. He limited the number of *Agents de Change* to 60 individuals willing to pay a price for the privilege and to post bond with the government for claims made by disappointed customers.

The large number of traders on the London Stock Exchange made it very difficult to reach collusive agreements. Further, the membership from the beginning was divided into two groups: jobbers, who held inventories of the most widely traded securities and traded on their own account as principals; and brokers, who were not supposed to hold inventories but only to act as agents for customers outside the exchange. Their respective sources of earnings, bid-ask spreads versus commissions, made it difficult for the two groups to agree on any change in rules. The smaller number of traders in New York could and did collude to maintain both minimum commissions and a restricted number of high-volume securities. But they had constantly to deal with challenges from other exchanges in New York as well as competition from exchanges in other cities. Brokers in Paris supposedly under the control of the central government were strictly forbidden to act as principals. Their small numbers and long tenure enabled them to effectively influence a succession of govern-

ments in order to maintain their personal profits.

II. Incentives

In the London Stock Exchange the separation of ownership of the marketplace from its operation meant that two committees, the Committee of Trustees and Proprietors (representing owners) and the Committee for General Purposes (representing users), were jointly vested in ultimate control. The Proprietors wished to maximize the return on their investment; and since their profits were based on the dues paid by Members, they pushed for as large a membership as possible. The Members, in contrast, wanted to maximize the volume of business (their income depended on the commissions they charged), but they did not care where the sales were made. The Proprietors demanded that the Members be independent of any financial institution, so that all trades passed through the exchange rather than simply the net purchases and sales after customers' trades had been offset against each other. In addition they frowned on any limitation on membership, a limitation that would be a necessary condition for policy that established minimum commissions. The end result was a highly competitive marketplace.

With such a large and diverse number of members, the brokers were unable to enforce minimum commissions and, therefore, were unwilling to limit the number of listings. As the number of brokers increased more rapidly than the number of jobbers, the Committee for General Purposes reflected more and more the interests of the brokers. It determined that jobbers were supposed to quote both their bid and ask prices for a particular security to any broker who inquired without knowing whether the broker had a commission to buy or to sell or in what amount. In this way a broker could quickly find the best price available for his client, make the deal on the spot, and take his modest commission expeditiously. The difficulty with this ideal was that it did not work at all for the great majority of stocks listed on the London Stock Exchange by 1878. Those stocks were seldom traded, so dealers were not willing to take positions without knowing in advance how many shares were involved in

the broker's commission and whether it was for a buyer or for a seller. Dealers frequently insisted on knowing the full details of the brokers' commissions, while brokers tried to deal directly with one another, eliminating the jobber (British Parliamentary Papers, 1878 pp. 129–42). New trading opportunities simply increased the conflict between jobbers and brokers. A jobber could make direct contact with off-site brokers and earn substantial profits from the differences in prices that a security commanded in different places; however, his increased business came at the expense of on-site brokers (Morgan and Thomas, 1962 Ch. 9). By 1912, the Members, now largely brokers, voted to enforce minimum commissions, and to outlaw shunting of deals to outside brokers by the jobbers (Morgan and Thomas, 1962 p. 157).

As the membership continued to increase, the interests of the Proprietors and Members tended to converge. As early as the 1870's when the volume of business outgrew the existing physical facilities a new, much larger building had to be financed. The Proprietors initially proposed to raise both the entrance fee and the annual subscription, while the Members proposed a new issue of capital stock. In the end, the more numerous Members won out as changes in the Deed of Settlement in 1875 and 1882 increased the original 400 shares to 20,000 shares and stipulated that all new shareholders had to be Members (Morgan and Thomas, 1962 p. 144).

The forerunner of the New York Stock Exchange was initially organized in 1792 with the Buttonwood Tree Agreement, but a more formal organization was established a quarter-century later in 1817 when 28 brokers adopted a new constitution. It established a new organization, the New York Stock and Exchange Board; but it largely formalized the original agreement to maintain exclusive dealings and minimum commissions (E. C. Stedman and A. N. Easton, 1969 pp. 62–67). In 1863 the exchange officially became "The New York Stock Exchange," and by 1867 it had 200 members and annual business in excess of \$3 billion.

Despite these gains, the New York Stock Exchange still could not claim to be the nation's, or even New York City's, premier

exchange. Most transactions still took place on the curb, and even the Open Board, one of the competing exchanges, did more business (Robert Sobel, 1970 pp. 41–42). In 1869, New York City's financial structure was dramatically altered. The 533 members of the New York Stock Exchange joined with the 527 members of the Open and Gold Boards, and the modern New York Stock Exchange was born. Memberships also became salable, and any new member had to purchase a seat from a retiring member. Between 1875 and 1909 the number of shares traded increased more than five times, and the real value of state and railroad bonds traded more than doubled. By merging with the primary exchanges dealing in nongovernment securities, the revitalized New York Stock Exchange offset the continued decline in the amount of government debt available for trade.

The path taken by the Paris Bourse to become the world's third-largest exchange by 1914 was strikingly different from either the London or the New York stock exchanges. From its beginning, it had the sanction of a government monopoly, a privilege its members gained by paying entry fees to the government and posting guaranty bonds. The government then maintained the building and controlled potential competitors by force. It also strictly limited the number of *Agents de Change* allowed to operate in the Bourse. The highest number ever allowed was 100 during the Napoleonic period, but this figure was reduced to 60 under the Restoration in 1816. It remained at that level until a series of reforms enacted at the end of the 19th century raised the number to 70 (Gustave Boissière, 1925 pp. 26–7, 33). Also, from its beginning, it attracted a fringe of interested but unsanctioned participants eager to find profits from the business conducted in the official marketplace; they operated in the legally unsanctioned *Coulisse*.

As a privileged company of the government, the Bourse relied on the government to locate and house it at an appropriate site within Paris. In 1826 it occupied the building next to the Palais Royal, but even after it had been expanded in the years 1901–1903, the *Compagnie des Agents de Change*, the proprietors of the *Parquet*, occupied their building solely as

renters from the city of Paris; the city had become the owner in 1829 (E. Vidal, 1910 pp. 118–19). That expansion, however, allowed the participants of the *Coulisse* to move into adjacent quarters and become assimilated more properly as complementary, rather than competing, agents to the official *Agents de Change*.

III. Implications for Regulation

Both the London and New York exchanges remained self-regulating organizations throughout this period. Periodic crises led to Parliamentary investigations, investigations that typically ended with minor pieces of legislation designed to placate the upper classes. Parliament's major acts served to enlarge the possible scope of trading activity for the London Stock Exchange. For example, it repealed the Bubble Act of 1720 in the middle of the crisis year of 1825. Then the Joint Stock Companies Act of 1844 encouraged the formation of joint stock companies in general, leading to the establishment of limited liability for joint stock corporations with Lowe's Act of 1856. True, some acts restricted speculative practices of one kind or another, but these were consistently ignored. The Members were more responsive to sanctions imposed by the Committee on General Purposes than to the possibility of losing lawsuits brought by outsiders. Only jobbers, who always acted as principals in the transactions, were really subject to the law governing enforcement of contracts.

In New York, stock-market panics also produced investigations, but only by the state legislature. The legislators in Albany were easily, and frequently, bribed into rescinding threatened regulations. The regulations of the New York Stock Exchange were, in fact, revised only in response to changes in the competitive threat of competition from other exchanges, whether in New York or elsewhere in the country. In the last decade of the century, the Exchange was able to institute two rule changes that strengthened the Exchange's imprimatur of quality but which competitive threats had previously prevented the Governing Committee from implementing. In 1892, after three failed attempts, the Governors finally established a clearing mechanism. By the

end of the century it encompassed almost all listed securities (Sobel, 1965 p. 131; John Grosvenor Wilson, 1969 pp. 423–32). In 1895 the Governing Committee voted to require that listed companies file annual reports, although it is clear that their word was still not law: they received no reports in either 1895 or 1896. By 1900, however, annual reports including both audited balance sheets and profit and loss statements became a prerequisite both for initial listing and for retaining that listing (Sobel, 1970 pp. 123, 177).

In France, the government's regulatory role varied with changes in political regime, but those changes usually affected the *Coulisse* more than the *Parquet*. The relative stability of the *Parquet*, in turn, can be attributed to the organizational strength of its *Compagnie*, composed as it was of a small number of individuals with life tenure. Its internal cohesion was strengthened further when in 1816 the government asked the remaining individual agents (their number had dwindled to 50 at the end of Napoleon's reign) to pay an additional 25,000 francs for their offices. In return the government made it possible for each *Agent de Change* to name his successor. Thus, although the government formally continued to control the nomination and the disposition of the title, the current titleholder now had a property right that could be sold. The *Agents de Change* were no longer civil servants named for life, but public officers possessing specific powers. The act of 1816 also strengthened the self-governance of the *Compagnie*, restoring a *Chambre Syndical* that enjoyed the triple powers of recruitment, discipline, and regulation. The corporate solidarity that naturally arose within the *Compagnie des Agents de Change* enabled them to exercise effective influence on the government to maintain the *Agents'* privileged position within France. The power of the Minister of Finance over the operation of the Bourse was effectively conceded to the *Compagnie*.

IV. Competing and Complementary Exchanges

By the 1830's London had preempted all competing exchanges in the metropolis. Thereafter, the only issue was the relationship with the regional exchanges. Jobbers in the

London Stock Exchange saw those exchanges as useful complements to their business: brokers in the regional exchanges expanded the jobbers customer base. Brokers saw them as annoying competitors, especially as the branch banking facilities of the joint stock banks enabled them to tap a countrywide customer base. As long as the jobbers maintained their effective check on the operating rules of the London Stock Exchange, the regional exchanges served as complements and helped increase the volume of business. When in 1912 the brokers finally enforced minimum commissions and abolished shunting, the regional exchanges became competitors. Indeed, the nationally integrated system of exchanges was broken up by this endogenous change in micro structure.

In New York, the Curb had existed somewhat uneasily alongside the New York Stock Exchange. Between 80 and 90 percent of its business was carried out on behalf of members of the formal exchange. Gradually, as the Curb became a recognized part of the evolving securities market, its relations with the New York Stock Exchange became better defined. In 1909 the representatives of the Exchange argued, "the curb market represents, first, securities that cannot be listed; second, securities in the process of evolution from reorganization certificates to a more solid status; and third, securities of corporations which have been unwilling to submit their figures and statistics to proper committees of the Stock Exchange" (New York State, 1909 p. 44). A listing on the New York Stock Exchange provided a substantial guarantee of stability; the Curb provided a market for riskier and more uncertain securities. The two had once again become complements rather than competitors.

In Paris, the *Coulisse* originally served a complementary function to the official *Parquet*. It provided counterparties to agents of the *Parquet* who were seeking matching buy or sell orders but were unable to serve as dealers themselves. They did, however, compete with the *Parquet*'s brokerage business; and in 1823, and again in 1850, the police were called in to remove them (Boissière, 1925 p. 142). The business of the *Coulisse* expanded rapidly under the Second Empire, and the *Coulistiers* formed two separate markets: one

dealing in government *rentes* and the other in securities not yet listed on the official exchange. The *Coulistiers* were mainly bankers dealing on behalf of the large joint stock banks and their customers. They occupied the outer hallways and colonnades of the Bourse before and after official trading hours and then regrouped to the lobby of the *Crédit Lyonnais* in the evening. In the 1890's, the volume of their business was more than 50-percent greater than that of the official market. Legislation in 1898, however, put them back in their place as *remisiers*, or shunters, for the *Agents de Change*. They ended as complements, rather than competitors (Boissière, 1925 pp. 142–49).

V. Conclusion

Despite constant monitoring of the successes and crises occurring in the other leading exchanges, the three leading exchanges of the world in the 19th century developed in quite different ways. These separate paths of development were set into motion by the original definition of property rights in the marketplaces. For example, banks and bankers were excluded from the London and Paris exchanges. They were, however, allowed to buy seats on the New York Stock Exchange and to participate in the complementary *Coulisse* exchange in Paris. Clearing facilities for improving the efficiency of transactions came first and were most highly developed in Paris, and they came last to London. The central bank played no direct role in the operation of either London or New York, but was vital for financing the fortnightly clearings in Paris. In short, even for the most highly developed marketplaces in the world at the height of finance capitalism in 1914, history mattered!

REFERENCES

- Boissière, Gustave. *La Compagnie des Agents de Change et le marché officiel à la Bourse de Paris*, 2nd Ed. Paris: Librairie Arthur Rousseau, 1925.
- British Parliamentary Papers. *Report from the Commissioners on the London Stock Exchange with minutes of evidence appendix index and analysis*, Vol. 19. 1878; re-

- printed, Shannon, Ireland: Irish University Press, 1969.
- Michie, Ranald C. *The London and New York stock exchanges, 1850-1914*. London: Allen and Unwin, 1987.
- _____. "Different in Name Only? The London Stock Exchange and Foreign Bourses, c. 1850-1914." *Business History*, January 1988, 30(1), pp. 46-68.
- Morgan, E. Victor and Thomas, W. A. *The stock exchange, its history and functions*. London: Elek, 1962.
- New York State. *Report of Governor Hughes' Committee on Speculation and Commodities*. Albany, NY: State of New York, 7 June 1909.
- Sobel, Robert. *The Big Board: A history of the New York Stock Market*. New York: Free Press, 1965.
- _____. *The curbstone brokers: The origins of the American Stock Exchange*. New York: Macmillan, 1970.
- Stedman, E. C. and Easton, A. N. "History of the New York Stock Exchange," in Edmund Clarence Stedman, ed., *The New York Stock Exchange: Its history, contribution to national prosperity, and its relation to American finance at the outset of the twentieth century*. New York: Greenwood, 1969, pp. 423-32.
- U.S. Department of Commerce, Bureau of the Census. *Historical statistics of the United States: Colonial times to 1970*. Washington, DC: U.S. Government Printing Office, 1975.
- Van Antwerp, W. C. *The stock exchange from within*. Garden City, New York: Doubleday, 1913.
- Vidal, E. *The history and methods of the Paris Bourse*. Washington, DC: U.S. Government Printing Office, 1910.
- Wilson, John Grosvenor. "The Stock Exchange Clearing House," in Edmund Clarence Stedman, ed., *The New York Stock Exchange: Its history, its contribution to national prosperity, and its relation to American finance at the outset of the twentieth century*. New York: Greenwood, 1969, pp. 423-32.

The Peace Dividend in Historical Perspective

By HUGH ROCKOFF*

After the collapse of communism in Eastern Europe and the Soviet Union it was anticipated that there would be a "peace dividend" in the United States. Aside from a decrease in defense spending, the term meant different things to different people. To some it meant cutting taxes, but a common interpretation, the one I will follow here, was that federal spending on "social needs" would increase. Since the Gulf War there have been substantial cuts in military spending, so it is natural to look to the experience with past demobilizations for insights into how the current cutbacks will affect the structure of the federal budget. History does seem to provide a lesson. A reading of (some) of the literature on the relationship between war and the federal budget might have led one to expect a substantial peace dividend.

I. The War-Ratchet Hypothesis

Wars, it is frequently argued, produce upward "ratchets" in federal spending.¹ One argument stresses the self-aggrandizing proclivities of the bureaucracy. Civilian agencies that grow during wars fight hard and successfully to transfer those resources to civilian uses afterwards. William Niskanen's (1974) model of government, although not specifically about wars, can be invoked as a motivation. A second argument stresses taxes. Taxes are raised during wars, people become reconciled to them, and so afterwards governments face lim-

ited political costs if taxes are reduced only part of the way to prewar levels.

The war-ratchet story appeals to both ends of the political spectrum. To conservatives it suggests that the growth of government is an historical artifact: opportunistic bureaucrats used wars to expand their domains. To liberals it suggests that the welfare state is held in check by exaggerated fears of higher taxes: after wars people recognize that higher taxes are not so bad.

There is, to be sure, a wide range of opinions about the war ratchet. There are skeptics such as Thomas E. Borcharding (1977 p. 37), who refers to it as that "insubstantial hypothesis"; R. A. Musgrave and Peggy B. Musgrave (1980 pp. 158-59), who conclude cautiously that the hypothesis "cannot be taken to give a conclusive explanation"; and Sam Peltzman (1980 p. 214), who expresses reservations based on international comparisons. But there are also writers who place great stock in the war ratchet: Herman Krooss (1966 p. 468) writes that "in each major war, government clearly became the pre-eminent player on the economic stage, and when the war was over, the government never completely reverted to the status of bit player";² Jack Hirshleifer (1976 p. 486) suggests that one explanation of the growth of government may be that wars leave in their wake a "mass of officeholders" who can resist cutbacks; and Jonathan Hughes (1991 p. 189) has given the war ratchet an exuberant endorsement. Perhaps the most influential defense of the war ratchet is Robert Higgs (1987). To be fair, Higgs is concerned more with "the growth of coercive power" (1987 p. 27) than with spending, and as much with peacetime crises such as the Great Depression as with wars. Nevertheless, Higgs's vigorous advocacy of the war-ratchet hypothesis has reinforced the belief that substantial

* Department of Economics, Rutgers University, New Brunswick, NJ 08901. I thank Michael Bordo, Hope Corman, Michael Edelstein, Stanley Engerman, Richard Sylla, Paul Trescott, and Eugene White for a number of useful comments on an earlier draft. Paul Trescott graciously shared his unpublished work sheets.

¹ Ratchet theorists have included a wide range of issues (e.g., the size of the federal labor force, the regulation of economic activity, moral and intellectual leadership, and so on) within their story. Here I focus on spending, because this is at the heart of the debate over a "peace dividend."

² Quoted in Claudia D. Goldin (1980 p. 948).

war ratchets in civilian spending are typical. Dwight R. Lee and Richard K. Vedder (1996), for example, interpret Higgs in this way.

The reason for seemingly contradictory conclusions derives, I believe, from a failure to distinguish consistently between the cost of past wars and the costs of purely civilian programs. Wars entail substantial costs that must be repaid over a long period of time. The most important cost, quantitatively, is interest on wartime debts. At one time, economists opposed wartime borrowing; now they recognize that borrowing by smoothing tax rates over time maximizes output over the long run. Veterans' benefits are next in importance, and until the Great Depression these constituted a significant fraction of the federal budget. In addition, there are numerous other costs, including reparations, relief in war-torn areas, and compensation for commandeered assets. In some cases the distinction between civilian spending and the cost of past wars is unclear. World War II, for example, produced the atomic bomb and the Atomic Energy Commission. Were the expenditures of the Commission costs entailed by the war, or were they a peace dividend? In general, however, few cases fall in the gray area.

Table 1 highlights this distinction. Spending is divided into three categories: (i) current military spending, (ii) cost of past wars, and (iii) civilian spending. Each is shown as a percentage of GNP before, during, and after each war. In some cases a transition year or years have been shown.

Similar tables have been drawn up before by M. S. Kendrick (1955), Goldin (1980), Jacob Metzger (1985), and Lee and Vedder (1996), among others. My table is not radically different, but I have been able to utilize some series that became available after Kendrick's (1955) pioneering work, and I have spent a bit of time ferreting out some of the indirect costs entailed by wars, such as relief in war-torn areas, that are usually neglected.

Two conclusions are obvious from the table. First, there has been a war ratchet in the sense that total spending has increased across nearly every war. But second, the reason for this is that the cost of paying for past wars and, in most cases, military spending remained higher

TABLE 1—THE IMPACT OF WARS ON FEDERAL SPENDING
(PERCENTAGE OF GNP)

Period	Current military spending	Cost of past wars	All other (civilian) spending	Total
War of 1812				
Prewar, 1807–1811	0.72	0.52	0.27	1.51
War, 1812–1814	2.78	0.44	0.25	3.47
Transition, 1815–1816	2.12	0.65	0.33	3.10
Postwar, 1817–1821	0.95	0.74	0.36	2.06
Mexican War				
Prewar, 1842–1846	0.93	0.16	0.43	1.53
War 1847–1848	1.56	0.27	0.33	2.16
Transition, 1849	0.91	0.51	0.46	1.87
Postwar, 1850–1854	0.67	0.30	0.76	1.74
Civil War				
Prewar, 1857–1861	1.71	0.12	0.78	2.61
War, 1862–1865	10.21	1.11	0.44	11.76
Transition, 1866	2.80	2.16	0.52	5.47
Postwar, 1867–1871	0.80	2.14	0.64	3.58
Spanish-American War				
Prewar, 1893–1897	0.47	1.28	0.84	2.59
War and transition, 1898–1899	1.26	1.13	0.86	3.24
Postwar, 1900–1904	0.83	0.83	0.81	2.47
World War I				
Prewar, 1914–1916	1.15	0.51	0.73	2.39
War, 1917–1919	15.48	1.11	0.63	17.21
Transition, 1920–1921	3.08	2.31	0.73	6.12
Postwar, 1922–1924	0.86	2.19	0.69	3.74
World War II				
Prewar, 1938–1940	2.16	1.52	6.94	10.62
Transition, 1941	12.73	1.29	5.30	19.31
War, 1942–1946	30.04	2.72	4.29	37.05
Transition, 1947	5.54	6.47	3.60	15.61
Postwar, 1948–1950	5.40	6.11	3.89	15.41
Korean War				
Prewar, 1948–1950	5.40	6.11	3.89	15.41
War, 1951–1953	12.16	3.58	3.44	19.17
Postwar, 1954–1956	10.43	2.90	4.34	17.67
Vietnam War				
Prewar and transition, 1963–1965	8.15	1.85	8.37	18.36
War, 1966–1970	8.67	1.95	9.26	19.88
Transition and postwar, 1971–1973	6.44	2.14	10.94	19.53

Notes and Sources: War of 1812 and Mexican War: spending, M. S. Kendrick (1955); GNP, Thomas S. Berry (1978). Cost of past wars (CPW) includes indemnities paid to Mexico. Civil War: spending, Paul B. Trescott (1966); GNP, Berry (1978). CPW includes the Freedman's Bureau. Spanish-American War: spending, Kendrick (1955); GNP, Nathan S. Balke and Robert J. Gordon (1989). World War I: spending, Kendrick (1955); GNP, Balke and Gordon (1989). CPW includes some foreign aid. World War II and the Korean War: spending, U.S. Bureau of the Census (1975 [series Y472, 473, 474, 476, 485]); GNP, U.S. Bureau of the Census (1975 [series F1]). CPW includes some foreign aid. Vietnam War: various issues of the *Statistical Abstract of the United States*. Fiscal-year data are adjusted to calendar years. A more detailed description of sources is available from the author upon request.

after wars. There has been no general tendency for civilian spending as a share of GNP to increase across wars; there have been war ratchets, but not peace dividends.

There is evidence in the 19th century of two ratchets in military spending. Military spending rose after the War of 1812 as a result of the deficiencies in the American military establishment that were revealed by the war; and after the Spanish-American War as a result, in part, of new international commitments.

The evidence for civilian ratchets, also, is mixed at best. There was an increase in the civilian spending ratio across the War of 1812, which according to Paul Studenski and Herman E. Krooss (1952 pp. 93–94) was partly due to a revival of Albert Gallatin's program of internal improvements as a result of increased revenues. Some of these projects, however, were undertaken to restore facilities damaged in the war (the White House and the Capitol were burned) and to increase access to territories in which Native Americans had been defeated. In addition there were heavy claims by the states for their contributions to the war effort. Unfortunately, I have not located a detailed breakdown of spending to determine how much of the ratchet can be explained by these costs. There was also a civilian ratchet across the Mexican War, although a policy of fiscal retrenchment lowered the ratio to its prewar level by the mid-1850's.

The civilian spending ratio fell across both the Civil War and the Spanish-American War, leaving the ratio little different at the end of the century from what it had been on the eve of the Civil War. The Civil War experience in particular argues strongly against the war ratchet as a major determinant of the civilian-spending ratio because the increases in taxes and spending were so large. There is a margin of potential error around the GNP estimates and, to a lesser extent, the spending estimates.³ One cannot be certain that there was no increase. Nevertheless, it is safe to conclude that war ratchets contributed little if anything to a long-run growth in the civilian spending ratio.

³ The table uses Trescott's (1966) estimates for the Civil War, which adjust for the premium on gold. Kendrick's estimates produce a small positive ratchet.

One might dismiss the lack of evidence for a civilian-spending ratchet in the 19th century with the argument that faith in *laissez-faire* was so ingrained that it was impossible to build a coalition of bureaucrats and politicians capable of overcoming resistance to increased spending (Higgs, 1987 pp. 79–84). There is little more evidence, however, for civilian-spending ratchets in the 20th century. The civilian ratio was about the same after World War I as it was before the war, a conclusion also reached in John Maurice Clark's (1970 p. 105) classic study. After World War II the civilian ratio fell, and after the Korean War it was only 12 percent higher, mainly as a result of the growth of transfer programs.

The case of World War II raises the question of whether deflating by realized GNP biases the results because nominal GNP was so much higher than expected after the war, an issue that also comes up on a lesser scale in some of the earlier wars. To some extent the use of transition years allows one to abstract from unexpected cyclical movements such as the recession after World War I and the boom after World War II. I experimented with projecting nominal income from past observations to get expected levels. Using expected levels does not significantly change the picture so long as the war years are included in the regression. But projections based on the war years fail to deal with the reality that many people feared a return to the depressed conditions of the 1930's. At least part of the fall in the measured civilian ratio across World War II, one must concede, may be the result of GNP being unexpectedly high. On the other hand there is an offset: much of the increase in civilian spending that occurred after World War II was produced by transfer programs launched during the Great Depression, rather than by new programs or by the expansion of existing programs which could be attributed to a war ratchet.

The civilian ratio was higher after the Vietnam War. But this outcome, most would agree, was produced by the independent decision to enlarge domestic anti-poverty programs, combined with nearly automatic growth in certain transfer programs. The war, if anything, put a damper on Great Society spending programs.

In the 20th century, as in the 19th century, there were ratchets in total spending. World War

II and especially the Korean War (and related international confrontations) produced substantial revisions in the public's desired level of peacetime military strength. The world wars, moreover, entailed costs that were repaid over a long period of time. But there is little evidence of major transfers of resources to civilian spending as a result of wartime demobilizations.

II. The Return to Normalcy

Inertia in taxes and spending favored the war ratchet. Typically, however, inertia was overcome by a desire to return to the way things were before the war, to return to normalcy, in President Harding's phrase that perfectly captured the mood of America in the wake of World War I. This should surprise no one. Spending programs are produced by long campaigns of persuasion and hard bargaining. The level at which they are funded depends on the strengths of the political interests that support them. If the relative strengths of those interests are not changed by the war, then why should the level of government spending change?

The Civil War might have proved an exception. Southerners were out of the Congress during the war, and the Democratic party was weakened for a generation. The result was a mass of legislation (e.g., bills relating to transcontinental railroads, a school for the deaf, a national banking act, etc.) that would not have passed before the war. Even in this case, however, there was no increase in the civilian spending ratio. Other wars, because they produced no fundamental realignments in politics, had less chance of producing fundamental changes in federal spending.

Until the Great Depression, returning to normalcy meant paying off wartime debts.⁴ The debt incurred during the war of 1812 was repaid in the 1830's, and substantial amounts of the debts run up during the Civil War and the Spanish-American War were also repaid. A good start was made on the World War I debt during the 1920's. After World War II, how-

ever, there was little enthusiasm for repaying the debt, a consequence, perhaps, of changed attitudes about national debts (Herschel I. Grossman, 1990).

III. The Long Run

Although the instantaneous war-ratchet hypothesis fails, a delayed and diminutive variant followed some of the wars. The policy of paying off the debt after 19th-century wars gradually reduced interest costs and produced budget surpluses that exceeded sinking fund requirements. By the time these structural surpluses emerged, the desire to return ever closer to prewar normalcy had waned, and the question of what to do about surpluses became a major political issue.

The example that is the best analogue for the current fiscal situation, assuming perhaps prematurely that we are now entering a period of budget surpluses, is the period of surpluses in the 1880's produced by the gradual repayment of the Civil War debt. By the 1880's enthusiasm for reductions in the debt beyond those mandated by the sinking fund had waned and were further inhibited by the requirement that the Treasury could not pay more than par. Some reductions in the internal taxes that had been imposed during the war were made, and some minor ones were eliminated; but the taxes on alcohol and tobacco had spawned their own lobbies and had become entrenched, and revenues from them continued to grow despite, if not because of, cuts in rates. Cuts in the tariff, the other important source of revenue, were also controversial. A variety of schemes for spending the surpluses, ranging from improved coastal defenses to the reclamation of Western deserts, were proposed. After considerable debate, the surpluses were used to expand veteran's benefits (Theda Skocpol, 1992 Ch. 2). Spending on veteran's benefits reached its 19th-century peak almost 30 years after the end of the war.

IV. Conclusion

The decrease in military spending produced by the end of the Cold War has not produced the

⁴ The decisions to return to the prewar price of gold after the War of 1812 and the Civil War provide another indication of the strong pull of normalcy.

immediate impact on the trend in domestic spending that many had anticipated. This might seem to contradict a well-established historical "fact" that civilian spending ratchets upward as a result of wars. But a close look at U.S. financial history demonstrates that recent experience is typical: after wars, civilian spending, measured relative to GNP, typically has returned to its prewar trend. The forces responsible for a war ratchet have been offset, more often than not, by the demand that the economy return to normalcy.

In the longer-run, however, one can identify a modest variant of the war-ratchet story that may be relevant in current circumstances. It has been conjectured that the United States is now entering a period of recurring surpluses in the federal budget. The main reasons for optimism have been unexpectedly strong growth in federal revenues, although limits on the growth of nondefense spending and cuts in military spending have played a role. Financial historians should be reminded of the 1880's when continuing budget surpluses were produced by internal taxes levied during the Civil War and falling interest payments. If the history of the 1880's is any guide, the current battle over who gets the surplus may prove to be as hard fought as the immediately preceding battle over who pays for the deficit.

REFERENCES

- Balke, Nathan S. and Gordon, Robert J. "The Estimation of Prewar Gross National Product: Methodology and New Evidence." *Journal of Political Economy*, February 1989, 97(1), pp. 38-92.
- Berry, Thomas Senior. "Revised Annual Estimates of American Gross National Product." Richmond, VA: Bostwick, 1978.
- Borcherding, Thomas E. "One Hundred Years of Public Spending," in Thomas E. Borcherding, ed., *Budgets and bureaucrats: The sources of government growth*. Durham, NC: Duke University Press, 1977, pp. 19-44.
- Clark, John Maurice. *The costs of the world war to the American people*. New York: Augustus M. Kelley, 1970 [originally published 1931].
- Goldin, Claudia D. "War," in Glenn Porter, ed., *Encyclopedia of American economic history: Studies of the principal movements and ideas*. New York: Scribner's 1980, pp. 935-57.
- Grossman, Herschel I. "The Political Economy of War Debt and Inflation," in William S. Haraf and Phillip Cagan, eds., *Monetary policy for a changing financial environment*. Washington, DC: AEI Press, 1990, pp. 166-81.
- Higgs, Robert. *Crisis and Leviathan: Critical episodes in the growth of American government*. New York: Oxford University Press, 1987.
- Hirshleifer, Jack. *Price theory and its applications*. Englewood Cliffs, NJ: Prentice-Hall, 1976.
- Hughes, Jonathan R. T. *The governmental habit redux: Economic controls from Colonial times to the present*. Princeton, NJ: Princeton University Press, 1991.
- Kendrick, M. Slade. "A Century and a Half of Federal Expenditures." National Bureau of Economic Research (New York) Occasional Paper 48, 1955.
- Krooss, Herman E. *American economic development*, 2nd Ed. Englewood Cliffs, NJ: Prentice-Hall, 1966.
- Lee, Dwight R. and Vedder, Richard K. "The Political Economy of the Peace Dividend." *Public Choice*, July 1996, 88(1-2), pp. 29-42.
- Metzer, Jacob. "How New Was the New Era? The Public Sector in the 1920s." *Journal of Economic History*, March 1985, 45(1), pp. 119-26.
- Musgrave, R. A. and Musgrave, Peggy B. *Public finance in theory and practice*, 3rd Ed. New York: McGraw-Hill, 1980.
- Niskanen, William A. *Bureaucracy and representative government*. Chicago: Aldine, 1974.
- Peltzman, Sam. "The Growth of Government." *Journal of Law and Economics*, October 1980, 23(2), pp. 209-87.
- Skocpol, Theda. *Protecting soldiers and mothers: The political origins of social policy in the United States*. Cambridge, MA: Harvard University Press, 1992.
- Studenski, Paul and Krooss, Herman E. *Financial history of the United States*. New York: McGraw-Hill, 1952.
- Trescott, Paul B. "Federal Government Receipts and Expenditures." *Journal of Economic History*, June 1966, 26(2), pp. 206-22.
- U.S. Bureau of the Census. *Historical statistics of the United States, Colonial times to 1970*. Washington, DC: U.S. Government Printing Office, 1975.

Wages and Labor Markets Before the Civil War

By ROBERT A. MARGO*

Documenting and explaining economic events in the past is the central task of economic history. This activity needs no more justification than any other type of "pure" economic research. But economic history can also provide useful perspective when events in the past share commonalities with events in the present, thereby rendering current economic change less mysterious or forbidding.

This paper describes a long-term research project aimed at documenting certain aspects of the economic history of labor in the United States before the Civil War. Although the labor history of the antebellum period seems remote from modern concerns, there are, perhaps surprisingly, numerous parallels—for example, changes in wage inequality, real-wage stagnation, high rates of unskilled immigration, the effects of labor-market conditions on welfare usage, wage rigidity and macroeconomic fluctuations, and "convergence" of geographically distinct labor markets.

I. Historical Issues

The antebellum period has always fascinated American economic historians. Between 1820 and 1860 America experienced the onset of industrialization and the spread of the factory system. Vast numbers of individuals moved to the frontier regions of the Midwest and South-Central states. Improvements in internal transportation and communications (canals, steamboats, railroads, the telegraph) produced enormous gains in internal commerce. In the late 1840's European immigrants arrived on

American shores in increasing numbers, permanently altering the economic, political, and social landscape. Slavery profited and proliferated in the South, reaching epic proportions as a political and moral dilemma that would only be resolved through a bloody, protracted, and extraordinarily costly Civil War.

The price of labor—more generally, how labor markets functioned—figures prominently in historical analysis of the antebellum economy. It has been said, for example, that common labor was relatively expensive, prompting more capital-intensive manufacturing than in England (H. J. Habbakuk, 1962). As economic development progressed after 1820, wage inequality allegedly worsened, as evidenced by a rise in the ratio of skilled (artisanal) to unskilled wages (Jeffrey G. Williamson and Peter H. Lindert, 1980). The westward movement of the population in the North seems puzzling at first glance, because estimates of regional per capita income show higher levels of income in the Northeast compared with the Midwest.

Considerable uncertainty exists over the long-run rate of growth of per capita income between 1820 and 1860, not to mention cyclical fluctuations. Because there are few sources on which to base estimates of gross national product before 1840, economic historians have inferred per capita income growth from movements in underlying components. For example, because measured output per worker in agriculture around 1840 was less than labor productivity in the nonfarm sector, the shift of labor out of agriculture was associated with per capita income growth (Paul David, 1967). But whether the productivity gap was a symptom of an inefficient intersectoral allocation of labor is unclear (Williamson and Lindert, 1980; Jeremy Atack and Fred Bateman, 1991). More generally, the ability of the antebellum economy to reallocate labor efficiently in response to secular economic forces or to "localized" shocks needs further investigation.

* Department of Economics, Harvard University, Cambridge, MA 02138; Department of Economics, Vanderbilt University; and Research Associate, National Bureau of Economic Research. I am grateful to Stanley Engerman, and workshop participants at Dartmouth, Harvard, Yale, the NBER, and the Economic History Association for helpful comments. This research is supported by the National Science Foundation.

Based largely on per capita income estimates and scattered real-wage series, many economic historians believe that the standard of living of the American working class improved significantly between 1820 and 1860. But labor historians are more pessimistic about the fruits of antebellum economic expansion, believing that real wages grew slowly in the long run and fluctuated considerably in the short run (Sean Wilentz, 1984). Consistent with the pessimist view, recent work in anthropometric and demographic history suggests that nutritional status underwent a surprising decline after 1840, and mortality also worsened (Margo and Richard Steckel, 1983; John Komlos, 1987; Clayne Pope, 1992).

II. New Evidence on Antebellum Wages

Research on many of the issues just mentioned has been seriously hampered by a lack of suitable wage data. No counterpart to today's *Current Population Survey* existed during the antebellum period. Economic historians have had to make do with various late 19th-century wage surveys that contain retrospective information dating back before the Civil War, or with scattered real-wage series for specific locations. Unfortunately, the extant sources are severely limited in temporal, geographic, and occupational coverage.

The primary purpose of my project is to present new evidence on wages over the 1820–1860 period. The evidence is drawn from two previously untapped archival sources. The first source is the *Reports of Persons and Articles Hired*, a collection of payroll records of civilian employees of the U.S. Army. Quartermasters at the various army installations were required to maintain monthly payrolls documenting the pay of civilian workers. Civilians were routinely hired to perform a wide variety of tasks at the posts, many relating to the construction, maintenance, or management of facilities. Duplicates of surviving payrolls are stored at the National Archives in Washington, DC. A large sample has been drawn from the National Archives collection, comprising approximately 55,000 wage observations. The sample covers all parts of the country and a

wide variety of artisanal, unskilled, and white-collar occupations.

The second source is the manuscript *Census of Social Statistics* of 1850 and 1860. As part of its enumeration effort, the census collected information for minor civil divisions on average daily, weekly, or monthly wages, with and without board, for various occupations (e.g., farm labor and common labor), and on the weekly cost of board. Although economic historians have long relied on aggregate wage statistics compiled from this source (Stanley Lebergott, 1964; Williamson and Lindert, 1980), the manuscripts have been little used. For the purpose of this project, I have collected manuscript data for 16 states, although most of the work is based on a smaller, eight-state sample, two from each of the principal census regions. All data from the census manuscripts are aggregated to the county level.

Use of data of military origin invites speculation about representativeness or sample-selection biases. In the case of the *Reports*, an obvious issue is whether the army systematically deviated from market wages in paying its civilian workers. Careful comparisons between wages paid to civilian workers at specific forts and wages paid by civilian employers in nearby locations suggest that the army simply paid the “going wage” in the local labor market (Margo, 1999 Ch. 2).

Although the *Reports* sample is very large by historical standards, it is not large enough to compute annual time series of average wages for narrowly defined occupations at specific locations (e.g., master carpenters at Jefferson Barracks, near St. Louis). Few forts operated continuously over the period, and none hired every type of labor in every year. Fortunately, enough information exists in the payrolls to estimate hedonic regressions. Separate regressions are estimated for three occupational groups (unskilled labor, artisans, and white-collar workers) for four census regions (Northeast, Midwest, South-Atlantic, and South-Central states). The data are pooled across forts within regions, and the regressions include dummy variables for the time period to which the observation pertains. The regressions fit the data well, and the patterns revealed in them (e.g., in wage differences between master artisans and apprentices) are

consistent with what is known from other sources.

Because the dependent variable (the nominal daily wage, or per diem equivalent) is expressed in log form, it is straightforward to compute nominal-wage indexes from the coefficients of the time-period dummies. To convert the nominal indexes into real form, I deflate by regional price indexes newly computed from wholesale price data originally collected by Arthur Cole (1938) and his associates. The problems with these price-deflators are such as to make today's debate over the CPI seem second-order. For example, the "Transportation Revolution" (George R. Taylor, 1951) lowered the costs of obtaining retail goods in rural areas, a trend missed in my deflators, as are various productivity advances that improved the quality of finished goods. The indexes omit the price of housing (by necessity, because the Cole collection has no data on housing prices), even though there is evidence that the relative price of housing, particularly in urban areas, rose between 1830 and 1860 (Margo, 1996). Sensitivity analysis indicates that these two omissions arguably offset each other (Margo, 1999 Ch. 4).

Table 1 shows estimates of "long-run" rates of growth of real wages, derived as the coefficient on trend in a linear regression. The rates of growth are mostly positive but (in the case of common labor and artisans) are generally lower than rates observed during the 20th century. Note that the growth rates for artisans do not exceed growth rates for common labor, inconsistent with previous claims that a "surge" in the skilled-wage premium occurred before the Civil War (Williamson and Lindert, 1980). However, the debate over the surge hypothesis has largely bypassed white-collar labor. Evidently the premium for white-collar labor did surge, perhaps the first such episode of a rising return to "educated" labor in American history. Because the surge was geographically widespread, a plausible candidate to explain it is general growth in the relative demand for managerial skills associated with the spread of commercial activity in the nonfarm sector.

Although the long-run trends were mostly positive, the new series also reveal that real-wage growth was highly variable before the

TABLE 1—LONG-RUN GROWTH RATES OF REAL WAGES, 1820–1860 (PERCENT PER ANNUM)

Region	Common labor	Artisan	White collar
Northeast	1.26	1.11	1.57
Midwest	0.72	−0.03	0.80
South Atlantic	0.75	0.24	1.27
South Central	0.83	0.57	1.56

Note: Growth rate is coefficient (β) of time trend in regression of log real wage: $\ln w = \alpha + \beta T + \varepsilon$.

Source: Margo (1999 Ch. 3).

Civil War. Generally speaking, wage growth was relatively slow in the 1820's and 1830's, rapid in the 1840's, and then stagnated again in the 1850's. At annual frequencies, some of the variability may be an artifact of excessive volatility in the price deflator, but the timing of prolonged fluctuations suggests that both nominal and real shocks played a role.

The conventional wisdom is that nominal wages were highly flexible before the Civil War (Peter Temin, 1969). Compared with labor markets today, the antebellum labor market was more like a textbook "spot" market. If by "flexible" one means that nominal wages responded to *sustained* changes in the price level, then antebellum wages were clearly flexible (Claudia Goldin and Margo, 1992). Still, nominal wages tended to lag behind changes in prices, both up and especially down. Some of this lag, however, may have been a consequence of real shocks that happened to be correlated with monetary shocks.

The late 1840's and early 1850's are a case in point. The discovery of gold in California fueled general inflation, while poor harvests in Europe drove up world food prices. A Midwestern railroad boom led to (region-specific) surges and then slumps in labor demand. From 1844 to 1856 the annual number of unskilled immigrants entering the country was 950-percent higher than the average during the 1830's; for skilled artisans, the corresponding figure was 279 percent (computed from U.S. Bureau of the Census, 1975 p. 111). In light of the immigration shock, it is not surprising that real wages, especially of the unskilled, stagnated during the 1850's.

The 1850's also witnessed the first "welfare explosion" in American history. The 1850 and 1860 *Census of Social Statistics* were the first to record the incidence of "poor relief," that is, the number of individuals receiving public assistance. Between 1850 and 1860 the relief rate rose by 76 percent (from 5.8 per 1,000 population to 10.2 per 1,000). As the antebellum economy industrialized, wage labor became more widespread, and workers became more dependent on market-generated incomes. Lacking the savings to smooth consumption, low-wage workers were especially vulnerable when, as in the 1850's, the price of food rose relative to unskilled wages. Using data from the 1850 and 1860 *Census of Social Statistics*, a simple "supply-demand" model of poor relief is estimated (Lynne Kiesling and Margo, 1997). Areas where the cost of food rose relative to unskilled wages between 1850 and 1860 experienced a substantial increase in the demand for public assistance, usually short-term. On the supply side, however, there was a steep negative trade-off between the relief rate and the average generosity of relief (measured by the amount spent per full-time-equivalent recipient). When faced with the sharp increase in demand, antebellum taxpayers were unwilling to maintain the generosity of relief at existing levels. Generosity fell, and consequently, the observed increase in relief was less than the increase in the demand for relief.

III. Labor-Market Integration

Well-functioning factor markets are crucial ingredients in long-term economic development. I use the data from the *Reports* and the *Census of Social Statistics* to examine various aspects of labor-market integration in the United States before the Civil War. The first aspect is wage "gaps" between the farm and nonfarm sectors for workers of comparable skill. Such gaps existed in other historical economies and continue to exist in many less-developed countries today, and they are widely viewed as impediments to economic growth. The census sample sheds light on this issue for the antebellum United States, because it contains information on wages paid to common nonfarm labor as well as farm labor.

These data need some manipulation because the farm wages include board and pertain to labor hired on a monthly basis, while the wages of common labor pertain to labor hired on a daily basis. When the data are properly analyzed, I find little or no nominal wage gap, on average, at the county level (a proxy for the local labor market) in either 1850 or 1860. But when the data are weighted to reflect differences in the geographic distributions of farm and nonfarm workers, a significant nominal wage gap emerges. The wage gap diminishes substantially when geographic differences in the cost of board (a proxy for the cost of living) are taken into account. Evidently, employers of nonfarm labor must have enjoyed some productivity (or nonlabor cost) advantage when located in predominantly nonfarm areas (Kenneth Sokoloff, 1986).

The United States is a vast country, and the antebellum economy faced the daunting task of reallocating labor from the East Coast to the frontier (the Midwest and South-Central regions), when information networks were inferior and internal transportation very costly by modern standards. Some scholars believe that the impediments to spatial arbitrage were so severe as to leave spatial wage differentials for lengthy periods of time. Using the wage data from the census, however, I find substantial wage convergence in the 1850's: high (or low) real-wage areas in 1850 were much less likely to be high (or low) real-wage areas in 1860, consistent with a process of spatial arbitrage.

To investigate regional patterns, I compute a set of regional benchmark cost-of-living estimates for 1850 which, in conjunction with the new wage series, permit me to measure changes over time in regional real-wage differences (e.g., the real wage of common labor in the Midwest relative to the Northeast). Initially (the 1820's) real wages were higher on the frontier (the Midwest and South-Central states) compared with settled areas (the Northeast and South-Atlantic states).¹ In the North,

¹ Because real wages differed across regions, growth rates of aggregate national series differ from the regional series. I compute aggregate national series by weighting the regional series by occupational shares derived from

the regional wage gap fell as the Midwest's share of the Northern labor force increased, a process of wage convergence. But in the South, the regional wage gap did not narrow appreciably over time, and a wage gap between the South-Atlantic states and the Northeast also emerged. Thus, while internal migration during the antebellum period was "efficient" in that labor flowed from low-value to high-value locations, shifts in regional labor demand (away from the "Old South" states of the South-Atlantic region relative to the South-Central and Northeast states) also influenced the regional evolution of wages.

Filling up the frontier was mostly an orderly process, in which states at the periphery generally were added to the existing stock in line with distance from the East Coast. A notable exception was California, which bypassed territorial status to become a state in 1850, as a consequence of the Gold Rush. The discovery of gold was an unexpected shock to labor demand of tremendous size that prompted a substantial reallocation of labor to a largely uninhabited region. In addition to the sample of payrolls previously described, I have collected a sample pertaining to California forts. Because the army was present in California before and after the discovery of gold, it is possible to construct nominal wage indexes spanning the Gold Rush period; these can be converted into real-wage indexes using price data collected by Thomas Senior Berry (1984). Real wages rose very sharply during the initial Gold Rush and then declined as labor migrated to the region. However, even though the Gold Rush itself was a transitory event, real wages appear to have remained permanently higher. Capital poured into California, fueling the growth of manufacturing and agriculture. In addition, the agglomeration economies associated with the extremely rapid growth of San Francisco after 1848 may have

been an important factor keeping wages high. In effect, California "leapfrogged" over other territories in the process of frontier development.

IV. Concluding Remarks

The specific features of any historical economy—its endowments and "deep structural parameters"—are, by definition, anachronistic. Yet economic historians know that much of what passes for economic novelty has precedents in the past. Real-wage growth before the Civil War was sluggish or stagnant for extended periods, just as it has been during the past 25 years. Changes in labor demand and labor supply during the antebellum period altered the distribution of wages, as has happened in the United States since 1970. Antebellum development was aided by factor markets that allocated labor from low- to high-value locations, a process of obvious relevance to transition and less-developed economies today. Careful scrutiny of past economic change, such as during the antebellum period of American history, provides a context for interpreting economic change in later periods.

REFERENCES

- Atack, Jeremy and Bateman, Fred. "Did the US Industrialize Too Slowly?" Working paper, University of Illinois, 1991.
- Berry, Thomas Senior. *Early California: Gold, prices, trade*. Richmond, VA: Bostwick, 1984.
- Cole, Arthur. *Wholesale commodity prices in the United States, 1700–1861*. Cambridge, MA: Harvard University Press, 1938.
- David, Paul. "The Growth of Real Product in the United States before 1840: New Evidence, Controlled Conjectures." *Journal of Economic History*, June 1967, 27(2), pp. 151–95.
- Goldin, Claudia and Margo, Robert. "Wages, Prices, and Labor Markets Before the Civil War," in C. Goldin and H. Rockoff, eds., *Strategic factors in nineteenth century American economic history: A volume to honor Robert W. Fogel*. Chicago: University of Chicago Press, 1992, pp. 67–104.

the 1850 census applied to Thomas Weiss's (1992) decadal figures on the nonfarm labor force (1820–1860), linearly interpolated between census dates (see Margo, 1999 Ch. 6). Regressions of the aggregate series on trend yield annual growth rates of 1.0 percent for common labor, 0.63 for artisans, and 1.55 percent for white-collar workers.

- Habakkuk, H. J.** *American and British technology in the nineteenth century*. Cambridge: Cambridge University Press, 1962.
- Kiesling, Lynne and Margo, Robert A.** "Explaining the Rise in Ante-bellum Pauperism, 1850–1860: New Evidence." *Quarterly Review of Economics and Finance*, Summer 1997, 37(2), pp. 405–17.
- Komlos, John.** "The Height and Weight of West Point Cadets: Dietary Change in Ante-bellum America." *Journal of Economic History*, December 1987, 47(4), pp. 897–927.
- Lebergott, Stanley.** *Manpower in economic growth: The American record since 1800*. New York: McGraw-Hill, 1964.
- Margo, Robert A.** "The Rental Price of Housing in New York City, 1830–1860." *Journal of Economic History*, September 1996, 56(3), pp. 605–25.
- . *Wages and labor markets before the Civil War*. Chicago: University of Chicago Press, 1999 (forthcoming).
- Margo, Robert A. and Steckel, Richard.** "Heights of Native-Born Whites During the Ante-bellum Period." *Journal of Economic History*, March 1983, 43(1), pp. 167–74.
- Pope, Clayne.** "Adult Mortality in America before 1900: A View from Family Histories," in C. Goldin and H. Rockoff, eds., *Strategic factors in nineteenth century American economic history: A volume to honor Robert W. Fogel*. Chicago: University of Chicago Press, 1992, pp. 267–96.
- Sokoloff, Kenneth.** "Productivity Growth in Manufacturing During Early Industrialization: Evidence from the American Northeast, 1820–1860," in S. L. Engerman and R. Gallman, eds., *Long-term factors in American economic growth*, Studies in Income and Wealth, Vol. 51. Chicago: University of Chicago Press, 1986, pp. 679–792.
- Taylor, George R.** *The transportation revolution, 1815–1860*. New York: Rinehart, 1951.
- Temin, Peter.** *The Jacksonian economy*. New York: Norton, 1969.
- U.S. Bureau of the Census, Department of Commerce.** *Historical statistics of the United States*. Washington, DC: U.S. Government Printing Office, 1975.
- Weiss, Thomas.** "US Labor Force Estimates and Economic Growth, 1800–1860," in R. Gallman and J. Wallis, eds., *American economic growth and standards of living before the Civil War*. Chicago: University of Chicago Press, 1992, pp. 19–75.
- Wilentz, Sean.** *Chants democratic: New York City and the rise of the American working class, 1788–1850*. New York: Oxford University Press, 1984.
- Williamson, Jeffrey G. and Lindert, Peter H.** *American inequality: A macroeconomic history*. New York: Academic Press, 1980.

USEFUL MICROECONOMICS FROM BUSINESS HISTORY[†]

Representative Firm Analysis and the Character of Competition: Glimpses from the Great Depression

By DANIEL M. G. RAFF*

Business history is terra incognita to most economists. They may have encountered the works of Alfred Chandler, most notably his magisterial *Visible Hand*. But they may well have taken away from its massive bulk little more than a sense of the inevitable coming of giant, divisionalized enterprise with a corporate planning staff devoted to measurement and planning. This is a vision congenial to the way we teach undergraduates about how firms work. But it is far from the most nourishing food for thought that Chandler's book offers, and it does not reveal much about method.

More likely, and certainly worse, economists may have encountered company histories—traditional business histories—in airport bookstores or used book shops. This genre will probably have been positively off-putting. The typical example contains much admiring prose, not much analysis, very little comparison, and practically no explicit theorizing. The best works provide provocative food for thought; but most will suggest to economists that business history is a subject better suited to the

Department of Public Relations than to the Department of Economics.

Much more fruitful encounters are possible. Businesses are basic units in the markets economists study. The internal organization of firms has a history, and actual competition between firms has a history that is at least as complex and rich. In both of these and in related areas, the history of business has a great deal of stimulus to offer economists and even economic historians.

One approach to such encounters involves addressing the archival materials worked by historians but with economist questions. The categories and principles that organize such collections are very far from the ones graduate training in economics suggests, and there are considerable gains from trade to be had simply by raising questions. A number of interesting (and eminently teachable) examples of this approach can be found in the NBER economic business history conference volumes (Peter Temin [ed.], 1991; Naomi R. Lamoreaux and Raff [eds.], 1995; Naomi R. Lamoreaux et al. [eds.], 1998). In terms of the title of this session, this type of work represents finding the microeconomics we know to be useful in the history of business.

A second and different approach, to be sketched in this paper, involves addressing more conventional economist data with questions raised by archival and other contemporary sources. This approach has several virtues: it is less of an intellectual stretch for working economists; it can often be carried out with materials that can be found in university libraries; and it actually offers the promise of extracting useful microeconomics and empirical research programs from business history, rather than simply reading microeconomics, textbook version or otherwise, into the history.

[†] *Discussants:* Bengt Holmstrom, Massachusetts Institute of Technology; Sidney Winter, University of Pennsylvania.

* The Wharton School, University of Pennsylvania, Philadelphia, PA 19104-6370, and National Bureau of Economic Research. Sidney Winter and Bengt Holmstrom gave thought-provoking comments. This paper discusses a stream of research carried out by Timothy Bresnahan and myself under a generous grant from the National Science Foundation. Seed money support from the Harvard Business School, the National Bureau of Economic Research, and Stanford University is also gratefully acknowledged. The usual disclaimer applies.

I. The Presenting Puzzle of Heterogeneous Behavior

I will exemplify this second approach with material on the organization of competition in American manufacturing during the Great Depression of the 1930's. Most of what economists think they know about this sector in this period comes from industry-level data compiled at the time by the Federal Reserve Board and the Department of Labor. As economists know from the gross-flows literature (see Steven J. Davis et al. [1996] for a summary of the empirics), such aggregates can hide a great deal of interesting behavior. A business-history-laden study of establishment-level data from the Depression period shows that what is missed in representative-firm analysis can have great significance.

The data behind the analyses summarized below far predate the data sets on which the modern panel work is based. The historical data come directly from the manuscript returns to the Censuses of Manufacturers conducted in 1929, 1931, 1933, and 1935. For many industries, these Census manuscripts survived in original or microfilm form with essentially complete coverage. The data are therefore suitable for panelization and analysis using sophisticated statistical methods.

However, even relatively unsophisticated methods yield heterogeneity crying out for further study. As a first exercise, one might create diagrams showing the flows of establishments into and out from productive activity across the Census years. The aggregate data suggest that the flow, such as it was, would be entirely outward; and this is demonstrably not true in the micro data for important industries. There are also dramatic differences between industries in the extent to which shake-outs occur. Neither of these patterns is unintelligible in terms of standard theory, of course; but standard data by itself is not illuminating as to why the patterns are occurring.

Two examples show the power of mobilizing additional information. The first concerns automobile manufacturing (details can be found in Timothy F. Bresnahan and Raff [1991, 1998]). The second concerns blast-furnace operations (i.e., the production of pig iron; the details for this are in Amy L. Bertin

et al. [1996]). Both examples have the feature that the key additional information lies on the boundary between strictly economic history and a history of decision-making within the enterprise. In both cases, the full interpretation raises questions that reach far beyond the particular circumstances to issues of how to interpret conventional data and how to view the interaction of firms and the markets behind them.

II. The Example of Automobiles

Aggregate data on the American automobile industry from 1929 through 1935 show a large and prosperous industry suffering a tremendous contraction and a mild recovery. But this is far too simple.

A figure following the careers of individual establishments and plotting their flows into and out of production (see, e.g., fig. 1 of Bresnahan and Raff [1991]) reveals a much more complex pattern. Between the peak and trough years of 1929 and 1933, the total number of exiting firms is nearly half the initial total: there is clearly a shake-out. There is entry over this period of nearly 10 percent as well. And substantial numbers of plants are mothballed, only to return in later years.

Comparing any pair of years, one can make a table comparing the attributes of the firms that were active in both years to the attributes of the firms that left and the firms that came in. Table 1, showing the peak and trough years, is an example. The firms that left look very different from either of the two other cohorts. Much of the adjustment that the aggregate data tempt one to attribute to "representative" firms in fact occurred at a different margin, through exit. The Depression was not a transitory incident. Industry makeup changed.

The microeconomist's natural temptation is to interpret these interfirm differences in terms of the height of average-cost curves (i.e., survivorship). More insight comes from investigating, with fewer presuppositions, what work was done and how it was organized in the establishments in question. Access to archival materials, old photographs, actual output, and the like is very helpful in this, but access to a microfilm reader, Inter-Library Loan, and the

TABLE 1—CHANGING COMPOSITION OF MOTOR VEHICLES
ESTABLISHMENTS: REVENUE, VEHICLES, AND
EMPLOYMENT PER PLANT, 1929–1933

Plants	1929	1933
Number:		
Continuing	106	106
Closing/new ^a	105	16
Wage-earner months (WEM): ^b		
Continuing	20,599	10,538
Closing/new ^a	4,931	3,586
Nominal revenues:		
Continuing	\$27.24	\$9.86
Closing/new ^a	\$7.83	\$3.22
Revenues (1929 \$ million): ^c		
Continuing	\$27.24	\$13.22
Closing/new ^a	\$7.83	\$4.32
Vehicles: ^d		
Continuing	36,564	16,465
Closing/new ^a	13,173	6,128
Nominal wage bill (\$ million):		
Continuing	\$2.83	\$0.94
Closing/new ^a	\$0.62	\$0.28
Wage bill (1929 \$ million): ^c		
Continuing	\$2.83	\$1.27
Closing/new ^a	\$0.62	\$0.37
Nominal salaries: ^e		
Continuing	\$629,626	\$231,839
Closing/new ^a	\$140,154	\$66,284
Salaries (1929 \$): ^{c,e}		
Continuing	\$629,626	\$310,870
Closing/new ^a	\$140,154	\$88,876
Total horsepower: ^f		
Continuing	7,101	8,874
Closing/new ^a	1,410	558
Vehicles/WEM:		
Continuing	2.09	1.50
Closing/new ^a	1.36	1.14
Revenues (1929 \$)/WEM: ^c		
Continuing	\$1,806	\$1,424
Closing/new ^a	\$1,244	\$1,055
Wage (wage bill/WEM) (1929 \$): ^c		
Continuing	\$132	\$127
Closing/new ^a	\$127	\$93

^a The number reported in the column for 1929 pertains to closing plants; the number in the column for 1933 is for newly opening plants.

^b This is the sum over 12 months of number of wage-earners employed.

^c The deflator is the predecessor to the CPI calculator in Robert A. Sayre (1948).

^d The number of vehicles produced includes chassis.

^e Of salaried employees.

^f Horsepower figures compare 1929 and 1935, because the census did not collect the information in 1933. The sample here is divided into plants present in 1929 and 1935, and those present in only one year.

Union List of Serials in Libraries of the United States and Canada (to identify trade journals and where they might be borrowed) would have been more than sufficient.

It emerges that in 1929, when the Depression struck, two distinct technologies were in place among the population of firms in the industry. Roughly speaking, all the establishments owned by the Big Three (General Motors, Ford, and Chrysler) deployed a clear precursor of the mass production known today. Some other establishments did as well. But many establishments deployed a technology that was to an important extent the earlier artisanal manufacturing technology. This technology relied to a significant extent upon skilled workers and judgment-intensive work.

Careful modeling of the exit decision reveals much more than that the simple survivorship model is wrong on the facts. It shows that the cost differences that were important were adjustment costs, not the costs of direct production behind the usual average cost curve. The technology-driven theory Bresnahan and I develop in our forthcoming paper places great weight on the costs that were treated by firms—as a matter of behavior, not of accounting—as fixed, and on the structure of sunkness of these fixed costs. This approach makes much more persuasive sense of the data and of what one can piece together from contemporary descriptions of products and production processes concerning skill requirements as well. The de facto fixed costs of the mass production plants may have been large, but they were mostly sunk. Even when these firms were faced with slack demand, not much expenditure could be saved. Thus, their plants had no great incentive to exit. This was not true in the relatively craft-intensive sector. The basic pattern is that those firms did exit. Faced with barriers to entry, they never returned.

Large composition shifts in the attributes of firm populations render aggregate data an unreliable guide to the behavior that economic theory actually models: the convenient abstraction of the representative firm disguises rather than clarifies what is to be explained. Shake-outs can thus be evolutionary events fundamentally altering the dynamic possibilities of the future, rather than mere features

behind a time series. The degree of sunkness of investment in industrial technology thus emerges as an interesting historical subject in itself.

III. The Example of Blast-Furnace Operations

Blast furnaces present a very different operation almost from the first. Their aggregates also show a tremendous shock at the coming of the Depression. But composition effects do not cause comparable interpretive problems: relationships in the aggregates are mirrored in the patterns of the establishment data. This is so despite very substantial interplant heterogeneity, the most common economic cause of nonaggregability.

Why is this so? Studying how the technology operated and what choices and incentives the owners actually faced again clarifies a great deal. Blast furnaces melt down raw iron ore in the presence of other chemicals in order to extract the iron. In the period in question, the output emerged from the furnace at temperatures in excess of 2,500° Fahrenheit (1,371°C), and subsequent stages of the production process required the iron to start at high temperatures. Competition between blast furnaces could not help but be very localized given positive transportation costs. The contrast to automobile manufacturing, in which each firm faced a national market and thus had many competitors, is striking. The output of blast furnaces producing pig iron of apparently identical chemical characteristics could be extremely poor substitutes.

The casual reader might conclude from the gross-flows literature that all industries are like automobiles in having large composition effects. The example of blast-furnace operations shows that this is not so and makes vivid a potentially much more general phenomenon. Transportation and communication have improved greatly over the intervening 60 years, and all else equal, this would tend to make the output of different plants more substitutable. But the relationship-specific investments at the center of so much theoretical research on the economics of competition in recent years can depress interplant substitutability in

precisely the same way as poor transportation and communication. The degree of competition between plants mediates between plant-level phenomena and aggregate outcomes. Understanding the evolution of this pattern, and its causes, is an important step in understanding the relationship between the history of the conduct of business and the aggregate economic history of the nation.

IV. Conclusions

For the purpose of economic analysis, traditional business histories give an excessively superficial look at firm-level decision-making. The approach I have illustrated in this paper represents carrying the logic of economists down to the choices managers actually confront as these present themselves. What makes such analysis business history is its focus on actors doing what business executives actually do at the office, which is making decisions. The objective of analysis is to understand the possibilities these actors see, the constraints they face, and the information they possess to guide their choices. It yields a picture of competitive dynamics (and, implicitly, of competitive strategy) and of the drivers of secular changes that is different from and very much richer than what comes naturally from the industry-level and even more aggregated data of the conventional statistical sources.

REFERENCES

- Bertin, Amy L.; Bresnahan, Timothy F. and Raff, Daniel M. G. "Localized Competition and the Aggregation of Plant-Level Increasing Returns: Blast Furnaces, 1929–1935." *Journal of Political Economy*, April 1996, 104(2), pp. 241–66.
- Bresnahan, Timothy F. and Raff, Daniel M. G. "Intra-industry Heterogeneity and the Great Depression: The American Motor Vehicles Industry, 1929–1935." *Journal of Economic History*, June 1991, 51(2), pp. 317–31.
- . "Plant Shutdown Behavior During the Great Depression and the Structure of the American Motor Vehicle Industry."

- Journal of Economic History*, 1998 (forthcoming).
- Chandler, Alfred D., Jr. *The visible hand: The managerial revolution in American business*. Cambridge, MA: Harvard University Press, 1977.
- Davis, Steven J.; Haltiwanger, John C. and Schuh, Scott. *Job creation and destruction*. Cambridge, MA: MIT Press, 1996.
- Lamoreaux, Naomi R. and Raff, Daniel M. G., eds. *Coordination and information: Historical essays on the organization of enterprise*. Chicago: University of Chicago Press, 1995.
- Lamoreaux, Naomi R.; Raff, Daniel M. G. and Temin, Peter, eds. *Learning by doing in firms, markets, and nations*. Chicago: University of Chicago Press, 1998 (forthcoming).
- Sayre, Robert A. *Consumers' prices, 1914-1948*. New York: National Industrial Conference Board, 1948.
- Temin, Peter, ed. *Inside the business enterprise: The use and transformation of information*. Chicago: University of Chicago Press, 1991.
- Union list of serials in libraries of the United States and Canada*. New York: Wilson, various years.

Survival and Size Mobility Among the World's Largest 100 Industrial Corporations, 1912–1995

By LESLIE HANNAH*

It is commonly observed that, among the largest corporations of today, there are some, like Du Pont, Siemens, or Shell, that were already global giants in the first decade of the century. Equally clearly, some giants of that era have declined to a shadow of their former selves. U.S. Steel was the largest industrial corporation in the world before World War I; if its then equity capitalization had matched the growth in the Standard and Poor's 500 index, it would still be worth much the same as today's largest industrial firm. Yet in fact, and notwithstanding its purchase of a giant oil company with twice its own equity in 1982, USX (as it is now rechristened) is only a tiny fraction of that size. Such divergent observations raise obvious questions: Which is the representative case? Does our collective memory play tricks, exaggerating the chance of survival? Has size mobility increased or decreased over time? These questions have generated studies of firm size mobility in individual countries over long time periods, and there have been some short-period studies of multinationals in the global economy. This paper begins to fill the obvious gap: it reports the initial findings on the survival and size mobility of the world's largest industrial corporations over the 20th century as a whole.

I. Approach and Data

The approach is based on extending backwards in time the listings published annually since 1987 by *Business Week*, though the population is limited to *industrial* (defined as manufacturing and mining) corporations. An exercise of this kind only makes sense if the global market is the relevant one. Factory and mine output has been extensively traded glob-

ally throughout the 20th century, even though in some mid-century decades that trade was temporarily subjected to extreme restraints. Hence, changes in the observed population of firms in these industries might reasonably be seen as the outcome of long-run evolutionary processes of global market selection. (By contrast, giant corporations in utilities, financial services, and other industries have been restricted by government regulations, national boundaries, or state ownership for most of the century.) *Business Week's* measure of corporate size, equity market capitalization, is readily available for most industrial companies over the century as a whole, though, to preserve comparability over time for the few industrial giants that were for substantial periods family-owned or state-owned, their "equity capitalization" was for some years proxied by balance-sheet assets (net of any quoted debt).

Table 1 summarizes the changes in the global giants by headquarters country between the terminal dates of the study: 1912 and 1995, both years that followed long periods of globalization (with strong international trade and capital flows) and international convergence (with high mobility of skills, techniques, and capital equipment between countries). The United States dominates the list at both dates. Although its share has fallen somewhat, this masks a substantial and sustained initial rise (with the U.S. share peaking at around four-fifths of global giants in the 1950's), reversed by a marked (but recently stemmed) fall. Germany experienced a sharper fall in its share, most of it concentrated in the earlier decades of the century, though smaller European states (presumably benefiting from European integration) have made up much of this deficit. The main gainer, with its gains concentrated in recent decades, is, of course, Japan. The "emerging markets" on the 1912 list (Mexico, South Africa, and Russia) have been a long time emerging. They have in fact re-

* City University Business School, Frobisher Crescent, Barbican Centre, London EC2Y 8HB, U.K.

TABLE 1—GEOGRAPHICAL DISTRIBUTION OF THE TOP 100 GLOBAL INDUSTRIAL FIRMS

Headquarters country	1912	1995
United States	52	40
United Kingdom	14.5	12.5
Germany	14	7
France	6	5
Small European countries	5.5	11.5
South Africa	4	1
Russia	3	0
Mexico	1	0
Japan	0	21
Australia	0	1
Korea	0	1
Total	100	100

Source: Adapted from Hannah (1998).

gressed: the largest Russian oil company of today (Lukoil) cannot yet match its Tsarist predecessor (Nobel Brothers) in relative size.

II. Findings

More generally, the giants at the top today are different firms than the giants of 1912. For example, only one of the five cigarette firms in the 1912 top 100 is still in the top 100 today (BAT Industries); contrariwise, today's leading global firm in tobacco manufacturing (Philip Morris) was in 1912 an insignificant tobacconist's shop. Nonetheless there is also surprising stability of the giants, not only in industries like petroleum (where it has been much remarked), but even where rapid change has been perceived. IBM, a favored current example of corporate hubris, actually entered the global top 100 several decades *before* it made its first electronic computer (it already had strong capabilities in mechanical business information processing globally in the 1920's), and its modern problems merely meant a fall from rank 1 to rank 9 by 1995: it remained in the top 100. In fact, 20 of the top 100 firms of 1912 were still in the top 100 of 1995: firms, like RTZ (Rio Tinto), Eastman Kodak, Guinness, Unilever (Lever Brothers), Procter and Gamble, Bayer (Elberfelder Farbenfabriken), BASF, GE, Exxon (Jersey Standard), and BP (Burmah/Anglo-Persian). The surviving giants were exclusively in "new" growth in-

dustries: petroleum, electricals, chemicals, copper mining, and branded products. Yet even in these industries, there have been significant long-run casualties. Cudahy Packing (one of the leading U.S. advertisers of branded products in 1912, but in terminal decline when it was acquired by General Host in 1968) has already been forgotten, and it may not be long before the same can be said of AEG (the "German GE" which was under bankruptcy protection in the 1980's, after consistent losses of market share to Siemens and ABB, and is now being dismantled by Daimler-Benz).

Perhaps surprisingly, the declining firms substantially outnumber the surviving giants in a population of firms defined *ex ante* by giant size in 1912, to avoid survivorship bias. A crude measure of their change in real size by 1995 is the increase in their market capitalization by that year (or on their earlier acquisition by another company), deflated by the Standard and Poor's 500 index. About two-thirds of this deflation is "accounted for" by general price inflation; the rest (about 2 percent per annum between 1912 and 1995) by general economic growth in which reasonably competent firms might have been expected to participate. Note that this is a very weak growth test. If a firm issued *any* new equity in the course of the century, for example, to finance acquisitions (as almost all these firms did), it could have been expected to do much better. By this generous measure, the average top-100 firm is 40-percent larger at the terminal date than in 1912, but that average is heavily influenced by a few firms like BP and Procter and Gamble (which were eight or nine times their 1912 size in 1995). The modal firm declined to zero size, and the median firm to 40 percent of its 1912 size. Of course the 1912 giant firms had greater probabilities of survival than small firms (whose probability of survival over the century approximates to zero), but in the long run they are subject to reasonably similar competitive processes. In the range between the average-sized joint-stock company of the time and the top hundred 1912 giants, it was necessary, in order to raise a firm's half-life (i.e., the period in which half the population disappears) by only one year, to increase its size by as much as 23 times (Hannah, 1998).

TABLE 2—NATIONAL CORPORATE PERFORMANCE
DIFFERENTIALS OF 1912 GIANTS, 1912–1995

Headquarters country	Percentages		
	United States	Germany	United Kingdom
Chances of survival in 1995 top 100	19	29	47
Chances of growth in size by terminal date	26	43	40
Average change in size by terminal date	+50	+20	+90

Source: Adapted from Hannah (1998).

III. Chandler's Hypothesis

This population of firms also makes it possible to undertake a test of Alfred D. Chandler's (1990) hypothesis that the British were markedly less successful than the Germans or Americans in establishing dominant firms in global oligopolies. This test avoids any survivorship bias of the kind that mars much generalization by business-strategy analysts. Table 2 shows the results for giant 1912 companies headquartered in these three countries (again using changing equity capitalization deflated by the Standard and Poor's index as a measure of growth). The data are inconsistent with Chandler's hypothesis (as indeed are wider samples and other observations [see Hannah, 1998]). Whatever generated Britain's poor 20th-century growth performance (it merely quadrupled its real GDP in this period, whereas America and Germany did twice as well as that), Britain's major global oligopolists are not the obvious culprits. More plausible causal factors suggested by others are perfectly consistent with these data: notably, the exceptional 20th-century performance of Germany's medium-sized companies (Gary Herrigel, 1996) and the remarkable and distinctive capacity of the United States to create innovation in Schumpeterian waves of creative destruction at the frontiers of 20th-century technology. As a comparison of Tables 1 and 2 indicates, while U.S. global giants of 1912 were most likely to disappear or decline, the United States nonetheless did better than Ger-

many and much the same as the United Kingdom in maintaining its overall share of the world's giants. The difference is accounted for by the United States' new giants, with no peer in other countries: the Coca-Colas, Boeings, Hewlett-Packards, and Cisco Systems of the modern global top 100. The point is eloquently made by tracking *Business Week's* July 1997 list of the global top-100 industrial corporations backwards to 1912: only 23 percent of today's U.S. giants were also in the 1912 top-100 list; the comparable figures for Germany and Britain are as high as 66 percent and 75 percent, respectively. Europe has been the home of Chandlerian stable oligopolies; America has more obviously been the creative (and destructive) dynamo of the Schumpeterian paradigm.

IV. Predictions

The comparison of successful and unsuccessful national giants over the long run also offers some pointers to likely future trends. One of the distinctive characteristics of the (unusually successful) British giants of 1912 was that they were unusually globally minded for their time. The British practiced free trade before World War I, while the American 1912 giants existed behind highly protective tariffs, and the German giants behind moderate ones. The British giants in this study thus more certainly began their 20th-century life as internationally competitive firms than their rivals, which may partly explain the surprising results in Table 2. Moreover, in 1912 German and American giant firms typically had only 10–15 percent of their assets or employment abroad, but British giants already had more than 30 percent of theirs abroad on average (Hannah, 1997) and thus were less constrained by home market performance. National differences in protectionism (though today perhaps mainly in nontariff barriers) and a fortiori in degree of multinationality or global-mindedness still exist, but they are less marked than in 1912. Today's top-100 giant industrial firms, wherever they are headquartered, clearly more nearly resemble the British firms of 1912 than the German or American ones.

If this generally acknowledged change translates into the generalization of the stability and growth rates of British global oligopolists shown in Table 2, then the global oligopolies of today may be expected to be much more stable in the next century than they have been in the present one. Of course, such a prediction flies in the face of the conventional corporate wisdom of recent years, which sees change accelerating rather than slowing down, though, in fact, no increase in turbulence is measurable over successive periods of the 20th century in the corporate population of this study. On the other hand, as Europe and Japan converge on the technological frontier previously uniquely set by the United States, they may adopt more of the characteristics of the Schumpeterian paradigm appropriate to that frontier position (though the cultural difficulties of that transition should not be underestimated). Opinions differ on whether stable oligopolies are conducive to the creation of new organizational capabilities or (in the antitrust tradition) of generating economic stag-

nation. If my prognostications of increased oligopoly stability are even half correct, there will be plenty of grist for industrial economists' mills in the 21st century.

REFERENCES

- Chandler, Alfred D., Jr.** *Scale and scope: The dynamics of industrial capitalism*. Cambridge, MA: Belknap, 1990.
- Hannah, Leslie.** "Multinationality: Size of Firm, Size of Country and History Dependence." *Business and Economic History*, Winter 1997, 25(2), pp. 144-54.
- . "Marshall's 'Trees' and the Global 'Forest': Were Giant Redwoods Different?" in Naomi Lamoreaux, Daniel Raff, and Peter Temin, eds., *Learning by doing in markets, firms, and countries*. Chicago: University of Chicago Press, 1998 (forthcoming).
- Herrigel, Gary.** *Industrial constructions: The sources of German industrial power*. New York: Cambridge University Press, 1996.

Partnerships, Corporations, and the Theory of the Firm

By NAOMI R. LAMOREAUX *

The purpose of this essay is to use business history, in particular the history of the contractual choices made by 19th-century entrepreneurs to organize their businesses, as raw material for reflecting on the nature of that entity we call the firm. The essay uses a transaction-costs framework to argue that no clear economic boundary distinguished ordinary contracts from those considered by law to be firms. Although my approach emphasizes the role of power and thus in some ways is similar to that of Oliver Hart (1995), Hart's analysis does not allow for important differences in the nature of ownership that resulted from the use of alternative organizational devices. By contrast, I argue that businesspeople could choose from a range of contractual forms that offered varying degrees of "firmness," that is, that differed in the extent to which they protected contracting parties against holdup. I also argue that increasing the degree of "firmness" of a contract was not always desirable from the standpoint of entrepreneurs.

The organizational choices available to entrepreneurs varied over the course of the 19th century. During the early part of the century, most businesses were organized as single proprietorships or partnerships. The only way to form a corporation was to secure a special charter from a state legislature, but such charters were usually granted only to projects deemed to be in the public interest. As the century progressed, however, the states gradually liberalized their policies on charters, and by the 1870's most had passed general incorporation laws that made the corporate form

widely available (Herbert Hovenkamp, 1991). Thus by the post-Civil War period, most businesses could organize either as partnerships or corporations.

The choice was an important one. On the most obvious level, partners had unlimited liability for their firm's debts, whereas the liability of members of corporations was limited. But there were other significant differences as well. For example, the partnership form of organization typically had a short time horizon. Partnership agreements either expired after fixed periods of time or included procedures for terminating the arrangement at the will of one of the members. The death of a partner also typically forced the dissolution of a firm. Although corporations could be organized for fixed periods of time, they were more commonly chartered in perpetuity. Moreover, the life of a corporation was independent of that of any of its stockholders. No one stockholder could force a dissolution of the firm, unless he or she owned a majority of the shares.

Another difference between partnerships and corporations, closely related to the issue of longevity, was the structure of governance. Members of a partnership all had the power to act as if they were the sole owners of the enterprise. As long as they were acting within the scope of the firm's normal business, they could enter into obligations that were binding on the firm without the consent of the other partners. Although partnership agreements might impose a hierarchical order on the firm's members and limit the actions that any one partner could take, such agreements did not exempt firms from liabilities assumed by partners contrary to their terms (Lamoreaux, 1995). Members of corporations, on the other hand, did not individually possess authority to bind the firm. The corporate form of organization concentrated management in the hands of officers elected by a vote of the stockholders, and the firm was not liable for debts incurred by members who were not empowered

* Departments of Economics and History, University of California, Los Angeles, CA 90095, and NBER. Research for this project was supported by fellowships from the National Endowment for the Humanities and the American Council of Learned Societies. I have benefited greatly from the comments of Bengt Holmstrom, Daniel M. G. Raff, Jean-Laurent Rosenthal, Peter Temin, and Sidney Winters.

to act on the stockholders' behalf (Edward H. Warren, 1929 pp. 21–22).

It is important to realize that these differences between the partnership and the corporation were not rigidly fixed, but rather varied in magnitude over time. Over the course of the century, as the courts worked through a wide variety of cases involving such matters as bankruptcies and stockholders' rights, they drew upon contending bodies of legal theory regarding corporations. During the early part of the century, the view that corporations were artificial creatures of the state held sway; by the middle it was increasingly common to view corporations as private contracts made by "aggregations" of businessmen; by the end, the courts were moving toward the view that corporations were legal persons in the eye of the law. During the period when aggregation theory dominated, the courts applied many aspects of partnership law to cases involving corporations, and the differences between the two forms narrowed (Morton J. Horwitz, 1992). During the same period, there were various circumstances in which stockholders of corporations might find themselves liable for amounts in excess of their investment. As Horwitz (1992 p. 94) has written, "the distinction between the liability of the members of a corporation and a partnership, so clear to modern eyes, was still regarded as a matter of degree rather than of kind."

It makes sense, therefore, to conceptualize the differences between partnerships and corporations, not in terms of discrete categories, but rather in terms of continuous variables that could take on different values at different points in time. In particular, the differences between these forms might be arrayed along two dimensions. The first dimension would be liability (the extent to which members of a firm were responsible for the enterprise's debts), with partnerships generally high on liability compared to corporations. The second might be thought of as a measure of the firm's autonomy (the extent to which it had a legal existence beyond that of its members). The ability of a partner acting alone to bind a firm and to dissolve it at will meant that partnerships were low on autonomy compared to corporations. These two dimensions were not, of course, completely independent, but they can

be distinguished analytically, and as I will show, it is useful to do so.

The literature on the advantages of the corporate form has focused almost exclusively on the first of these two dimensions, liability. Some scholars have argued that there were no clear benefits to limited liability because whatever savings it permitted in raising equity capital were offset by higher costs in securing loans. Others have countered that limited liability played an important role in lowering transaction costs, for example, those associated with transferring shares on the securities markets.¹ Arguments like the latter imply, however, that the benefits of limited liability were greater for large firms seeking public ownership than for small closely held companies. Although the historical record indicates that large enterprises almost universally adopted the corporate form, the vast majority of businesses choosing to incorporate during the late 19th century were small firms whose stock was closely held. It is important to understand why small as well as large firms made this decision, for the widespread adoption of the corporate form by small firms played a major role in its subsequent history.

In the case of small firms, there appears to be abundant evidence that liability rules in and of themselves did not determine the choice of organizational form. First, firms did not always take advantage of limited liability when it was readily available to them. Many British joint-stock companies, for example, could have obtained limited liability by reorganizing under a general incorporation law passed during the mid-1850's but chose instead to remain as they were (Forbes, 1986; Smart, 1996). Similarly, in the early 19th-century United States, small firms sometimes voluntarily wrote into their corporate charters clauses that specified unlimited liability. Throughout the century, moreover, it was common for the officers and leading stockholders of small corporations to endorse personally their company's debts in order to secure commercial credit and bank loans (Lamoreaux, 1997).

¹ For a survey of this literature, see Kevin F. Forbes (1986) and Michael Smart (1996).

A second type of evidence, which is perhaps more revealing about the determinants of contractual choice, was small firms' reluctance to adopt an available alternative: the limited partnership. Legislation permitting this type of organization was first passed in New York and Connecticut in 1822 and then adopted by most other states over the next couple of decades—long before the advent of general incorporation laws made the corporate form a widely available alternative—and yet few firms seem to have chosen the new form. This lack of interest is somewhat surprising, because limited partnerships would seem to have had some advantages over ordinary partnerships. The statutes created firms with two types of partners: general partners, who had unlimited liability and whose rights and responsibilities were the same as those of members of ordinary partnerships; and special partners, whose liabilities were limited to their investments and who had no authority over the management of the company.² From the standpoint of a general partner, the limited partnership functioned much like a limited-liability corporation in which officers assumed personal responsibility for debts in excess of the firm's capital. From the standpoint of a special partner (i.e., an investor who did not intend to participate in the management of the firm) the form involved the possibility of greater gains than could be obtained from a simple loan contract without the risks that an ordinary partnership entailed. Compared to a corporation, moreover, the arrangement may have reduced principal-agent problems because the firm's general partners not only shared fully in whatever gains or losses their actions generated, but were personally liable for debts in excess of capital. Moreover, investors who were dissatisfied

with the quality of the firm's general partners could vote with their feet and refuse to renew the partnership when its term expired. Stockholders of corporations could sell out only if there were willing buyers for their shares; limited partners could force the firm to reimburse them.

This, of course, was the crux of the problem from the standpoint of the general partners. If one adopts an Oliver Williamson type of transaction-cost view of the firm, the reason is apparent. According to Williamson, whenever resources have a significantly greater value in combination than they do in alternative uses, there is a risk of holdup which can only be countered by bringing the assets together within a single firm (Williamson, 1985). In general, however, firms organized as partnerships had much less ability to protect members against holdup than did firms organized as corporations for the simple reason that partnerships were lower in autonomy. The ability of partners to dissolve the firm at will or after the fixed term of the agreement meant that there were circumstances under which one party could use the threat of dissolution to force the others to grant more favorable terms. For example, during the first decade of the 19th century, E. I. Dupont's refusal to allow his partner, Peter Bauduy, to count as capital a note he had endorsed for the enterprise produced much "animadversion on the part of P. Bauduy, who threatened to sue for a dissolution of partnership to stop the factory and could not be pacified, but by the new agreement" in which Bauduy "exact[ed] from the concern some extra compensation and advantages."³

Ordinary partnerships were nonetheless an improvement over nonfirm contractual agreements, because they raised the costs that had to be borne by a party that failed at an attempt at holdup. If, instead of better contract terms,

² There were problems with the way the legislation was written which, by exposing special partners to unlimited liability under certain circumstances beyond their control, reduced the attractiveness of the form (William Draper Lewis, 1917). What is intriguing, however, is how little interest there was in remedying these statutory deficiencies. Eventually a few Western states passed legislation protecting special partners who acted in good faith, but that was all until the drafting of the Uniform Limited Partnership Act in the second decade of the 20th century (Warren, 1929 pp. 309–10).

³ "Answer of Eleuthère Irénée Dupont made in his own name as well as in behalf of Mess. E. I. Dupont de Nemours & Co. to the bill filed in chancery by Peter Bauduy against him and the said concern," 1817, Special Papers, Bauduy Lawsuit (Part I) (1805–1828), Longwood Mss. Box 45, Accession Group 5, E. I. du Pont de Nemours & Co., Series C, Hagley Library Manuscript Collections, Wilmington, DE.

the attempt resulted in the dissolution of the firm and the liquidation of its assets, all of the partners stood to lose from the forced sale of resources whose value outside the firm was less than it was within. In the case of limited partnerships, however, this principle of equal pain did not hold. Instead, the claims of the special partners took precedence over those of the general partners. At the end of the term of the limited partnership, the general partners either had to meet the special partners' conditions for renewal or face dissolution. In effect, they were in the position of borrowers who owed a big balloon payment at the end of a fixed period of time. If they could not get another loan, they would have to liquidate assets to pay off their debt. The risk of holdup, moreover, was not borne only by one side. If the firm was unusually successful, the general partners could extract higher payments from the special partners in return for their agreed-upon share of the profits. For example, Aaron Benedict reorganized his Connecticut button manufactory as a limited partnership in 1829. The firm did very well, and when the agreement expired in 1834, Benedict raised the price of the smallest share from \$1,000 to \$2,500. In 1838 he raised the price again to \$5,000. In both 1834 and 1838, small investors who could not come up with the requisite sums were forced to drop out of the company (Matthew W. Roth, 1994).

In other words, the limited partnership, although lower on the liability dimension than the ordinary partnership, was also lower on the dimension of autonomy. The latter seems to have outweighed the former in the eyes of 19th-century entrepreneurs, and few seem to have organized their enterprises as limited partnerships.⁴ The choice between the partnership and the corporate form was very different, however, for corporations were not only lower

on the liability dimension, but higher on the autonomy dimension compared to partnerships; that is, corporations also offered greater protection against holdup. The following example helps to explain why.

The case involves the company formed by George Corliss to manufacture the famous steam engine that bore his name. The firm was first organized as a partnership in 1847 and then, ten years later, reorganized as a corporation, with the bulk of the stock evenly divided between Corliss himself and an investor named Edwin J. Nightingale, who served as the firm's treasurer. The two men agreed to admit several additional members to the firm by selling equal amounts of their stock to the new parties, but when Corliss sought to increase the number of shares owned by his brother William, Nightingale refused to reduce his own holdings for this purpose. Perhaps he realized that the two brothers would likely vote as a block and that by selling stock to William he would in effect be giving George control of the company. After repeated attempts to persuade Nightingale to change his mind, George sold William a block of his own stock. But he also determined to force Nightingale to sell out his interest. If the firm had still been organized as a partnership, this action would have been relatively easy to take. Corliss could have unilaterally forced a dissolution of the firm and, at the same time, prevented a mutually disastrous liquidation by offering Nightingale cash for his share. The corporate form of organization made things much more difficult, however, because the two parties had to agree to a dissolution. In this particular case, Corliss had an ace up his sleeve. He retained personal control of the patents on his steam engine and used them to threaten to destroy the company by licensing competitors to make his patented engines.⁵ Nightingale was forced to capitulate, but later investors in similar situations would protect

⁴ In France, the limited partnership (known as *société en commandite*) was an important and commonly used business form. It had greater autonomy than the American variant, however, because it could be organized for long periods of time (typically 99 years) and included contract terms that gave the firm a lifespan independent of that of its general partners (Charles E. Freedman, 1979). I am indebted to Lacey Plache for the information on contract periods.

⁵ "Business from 1847 to 1861," Box 4, Folder 6, George H. Corliss Papers, Ms. 80.3; 13 February, 12 March, 19 March, 20 March, 1 June, 8 June, 24 June, 18 August, 19 August, and 24 September 1863, William Corliss Diaries, Ms. 80.4, Brown University Library, Brown University, Providence, RI.

themselves against holdup by making the assignment of full patent rights a condition of their participation in the corporation.

If one adopts Williamson's logic and thinks about firms as organizational arrangements that reduced opportunities for holdup, then the Corliss example suggests that it is not very useful to think about firms and nonfirms as either/or categories. Rather it is more useful (again) to think in terms of a continuum of contractual arrangements arrayed according to their degree of "firmness." Some contracts offered parties relatively little protection against holdup; that is, they had relatively little of that attribute I am calling "firmness." Others with more "firmness" offered stronger protection. In general, organizational arrangements that ranked higher on the autonomy dimension had more "firmness" than those that ranked lower.

The Corliss example also suggests that increasing the degree of "firmness" of an organization entailed costs for at least some of the contracting parties that were directly related to the greater protection against holdup. The relatively lower "firmness" of the partnership form of organization would have allowed Corliss to benefit from Nightingale's investment when he needed it and then to take control of the company when he did not. The relatively higher "firmness" of the corporation made this sequence of events much more difficult to effect. The general point I wish to make here is that use of the term "holdup" may have misleadingly negative connotations. What is really at stake is the ability of one party to a contract to use some form of economic muscle against another. Although the exercise of such muscle could have undesirable consequences for one or more parties, so could blocking its use. In other words, from the standpoint of the economic actors involved, there was a trade-off between greater protection against holdup and the ability to use economic power. The corporate form of organization was not always an improvement over the partnership for all of the parties concerned.

A couple of concluding observations relate the above discussion to the literature on the theory of the firm. First, the argument I have presented is similar to that of Hart (1995) in

the sense that it builds on the notion that incomplete contracts and power are keys to understanding the structure of economic institutions. For my purposes, however, Hart's notion of ownership, which he defines as the possession of residual rights of control, is not all that helpful because it is still basically an either/or proposition. It does not allow for subtle differences in the nature of ownership that result from employing different organizational forms. Second, my argument may also seem at first glance to be similar to that of theorists like Steven Cheung (1983), who argue that it is impossible to develop a workable definition of the firm that distinguishes it from other kinds of contractual arrangements. I would disagree, however. If one conceptualizes the central problem of economic development as the bringing together of producers and investors in a way that fosters the creation of sustained capabilities, then "firmness" becomes an attribute of great economic significance. Protection against holdup is crucial to actors' willingness to invest in the kinds of enterprise-specific capabilities that Richard Nelson and Sidney Winter (1982) have shown to be so important in economic progress. The question then becomes how much protection is optimal under different circumstances.

REFERENCES

- Cheung, Steven N. S. "The Contractual Nature of the Firm." *Journal of Law and Economics*, April 1983, 26(1), pp. 1-21.
- Forbes, Kevin F. "Limited Liability and the Development of the Business Corporation." *Journal of Law, Economics, and Organization*, Spring 1986, 2(1), pp. 163-77.
- Freedeman, Charles E. *Joint-stock enterprise in France, 1807-1867*. Chapel Hill: University of North Carolina Press, 1979.
- Hart, Oliver. *Firms, contracts, and financial structure*. New York: Oxford University Press, 1995.
- Horwitz, Morton J. *The transformation of American law, 1870-1960*. New York: Oxford University Press, 1992.
- Hovenkamp, Herbert. *Enterprise and American law, 1836-1937*. Cambridge, MA: Harvard University Press, 1991.

- Lamoreaux, Naomi R. "Constructing Firms: Partnerships and Alternative Contractual Arrangements in Early-Nineteenth-Century American Business." *Business and Economic History*, Winter 1995, 24(2), pp. 43-71.
- . "The Partnership Form of Organization: Its Popularity in Early-Nineteenth-Century Boston," in Conrad E. Wright and Kathryn P. Viens, eds., *Entrepreneurs: The Boston business community, 1750-1850*. Boston, MA: Massachusetts Historical Society, 1997, pp. 269-95.
- Lewis, William Draper. "The Uniform Limited Partnership Act." *University of Pennsylvania Law Review and American Law Register*, June 1917, 65(8), pp. 715-31.
- Nelson, Richard R. and Winter, Sidney G. *An evolutionary theory of economic change*. Cambridge, MA: Harvard University Press, 1982.
- Roth, Matthew W. *Platt Brothers and Company: Small business in American manufacturing*. Hanover, NH: University Press of New England, 1994.
- Smart, Michael. "On Limited Liability and the Development of Capital Markets: An Historical Analysis." Working paper, University of Toronto, 1996.
- Warren, Edward H. *Corporate advantages without incorporation*. New York: Baker, Voorhis, 1929.
- Williamson, Oliver E. *The economic institutions of capitalism*. New York: Free Press, 1985.

THE NEW INSTITUTIONAL ECONOMICS[†]

The New Institutional Economics

By RONALD COASE*

It is commonly said, and it may be true, that the new institutional economics started with my article, "The Nature of the Firm" (1937) with its explicit introduction of transaction costs into economic analysis. But it needs to be remembered that the source of a mighty river is a puny little stream and that it derives its strength from the tributaries that contribute to its bulk. So it is in this case. I am not thinking only of the contributions of other economists such as Oliver Williamson, Harold Demsetz, and Steven Cheung, important though they have been, but also of the work of our colleagues in law, anthropology, sociology, political science, sociobiology, and other disciplines.

The phrase, "the new institutional economics," was coined by Oliver Williamson. It was intended to differentiate the subject from the "old institutional economics." John R. Commons, Wesley Mitchell, and those associated with them were men of great intellectual stature, but they were anti-theoretical, and without a theory to bind together their collection of facts, they had very little that they were able to pass on. Certain it is that mainstream economics proceeded on its way without any significant change. And it continues to do so. I should explain that, when I speak of mainstream economics, I am referring to microeconomics. Whether my strictures apply also to macroeconomics I leave to others.

Mainstream economics, as one sees it in the journals and the textbooks and in the courses taught in economics departments has become more and more abstract over time, and although it purports otherwise, it is in fact little

concerned with what happens in the real world. Demsetz has given an explanation of why this has happened: economists since Adam Smith have devoted themselves to formalizing his doctrine of the invisible hand, the coordination of the economic system by the pricing system. It has been an impressive achievement. But, as Demsetz has explained, it is the analysis of a system of extreme decentralization. However, it has other flaws. Adam Smith also pointed out that we should be concerned with the flow of real goods and services over time—and with what determines their variety and magnitude. As it is, economists study how supply and demand determine prices but not with the factors that determine what goods and services are traded on markets and therefore are priced. It is a view disdainful of what happens in the real world, but it is one to which economists have become accustomed, and they live in their world without discomfort. The success of mainstream economics in spite of its defects is a tribute to the staying power of a theoretical underpinning, since mainstream economics is certainly strong on theory if weak on facts. Thus, for example, in the *Handbook of Industrial Organization*, Bengt Holmstrom and Jean Tirole (1989 p. 126), writing on "The Theory of the Firm," remark that "the evidence/theory ratio ... is currently very low in this field."

This disregard for what happens concretely in the real world is strengthened by the way economists think of their subject. In my youth, a very popular definition of economics was that provided by Lionel Robbins (1935 p. 15) in his book *An Essay on the Nature and Significance of Economic Science*: "Economics is the science which studies human behaviour as a relationship between ends and scarce means that have alternative uses." It is the study of human behavior as a relationship. These days economists are more likely to refer

[†] Roundtable discussion.

* University of Chicago Law School, 1111 East 60th Street, Chicago, IL 60637-2786.

to their subject as "the science of human choice" or they talk about "an economic approach." This is not a recent development. John Maynard Keynes said that the "Theory of Economics ... is a method rather than a doctrine, an apparatus of the mind, a technique of thinking, which helps the possessor to draw correct conclusions" (introduction in H. D. Henderson, 1922 p. v). Joan Robinson (1933 p. 1) says in the introduction to her book *The Economics of Imperfect Competition* that it "is presented to the analytical economist as a box of tools." What this comes down to is that economists think of themselves as having a box of tools but no subject matter. It reminds me of two lines from a modern poet (I forget the poem and the poet but the lines are indeed memorable):

I see the bridle and the bit all right
But where's the bloody horse?

I have expressed the same thought by saying that we study the circulation of the blood without a body.

In saying this I should not be thought to imply that these analytical tools are not extremely valuable. I am delighted when our colleagues in law use them to study the working of the legal system or when those in political science use them to study the working of the political system. My point is different. I think we should use these analytical tools to study the economic system. I think economists do have a subject matter: the study of the working of the economic system, a system in which we earn and spend our incomes. The welfare of a human society depends on the flow of goods and services, and this in turn depends on the productivity of the economic system. Adam Smith explained that the productivity of the economic system depends on specialization (he says the division of labor), but specialization is only possible if there is exchange—and the lower the costs of exchange (transaction costs if you will), the more specialization there will be and the greater the productivity of the system. But the costs of exchange depend on the institutions of a country: its legal system, its political system, its social system, its educational system, its culture, and so on. In effect it is the institutions that govern the performance of an econ-

omy, and it is this that gives the "new institutional economics" its importance for economists.

That such work is needed is made clear by another feature of economics. Apart from the formalization of the theory, the way we look at the working of the economic system has been extraordinarily static over the years. Economists often take pride in the fact that Charles Darwin came to his theory of evolution as a result of reading Thomas Malthus and Adam Smith. But contrast the developments in biology since Darwin with what has happened in economics since Adam Smith. Biology has been transformed. Biologists now have a detailed understanding of the complicated structures that govern the functioning of living organisms. I believe that one day we will have similar triumphs in economics. But it will not be easy. Even if we start with the relatively simple analysis of "The Nature of the Firm," discovering the factors that determine the relative costs of coordination by management within the firm or by transactions on the market is no simple task. However, this is not by any means the whole story. We cannot confine our analysis to what happens within a single firm. This is what I said in a lecture published in *Lives of the Laureates* (Coase, 1995 p. 245): "The costs of coordination within a firm and the level of transaction costs that it faces are affected by its ability to purchase inputs from other firms, and their ability to supply these inputs depends in part on their costs of coordination and the level of transaction costs that they face which are similarly affected by what these are in still other firms. What we are dealing with is a complex interrelated structure." Add to this the influence of the laws, of the social system, and of the culture, as well as the effects of technological changes such as the digital revolution with its dramatic fall in information costs (a major component of transaction costs), and you have a complicated set of interrelationships the nature of which will take much dedicated work over a long period to discover. But when this is done, all of economics will have become what we now call "the new institutional economics."

This change will not come about, in my view, as a result of a frontal assault on mainstream economics. It will come as a result of economists

in branches or subsections of economics adopting a different approach, as indeed is already happening. When the majority of economists have changed, mainstream economists will acknowledge the importance of examining the economic system in this way and will claim that they knew it all along.

REFERENCES

- Coase, Ronald H. "The Nature of the Firm." *Economica*, November 1937, 4, pp. 386-405.
- . "My Evolution as an Economist," in William Breit and Roger W. Spencer, eds., *Lives of the laureates*. Cambridge, MA: MIT Press, 1995, pp. 227-49.
- Henderson, H. D. *Supply and demand*. London: Nisbet, 1922.
- Holmstrom, Bengt and Tirole, Jean. "The Theory of the Firm," in Richard Schmalensee and Robert D. Willig, eds., *Handbook of industrial organization*. Amsterdam: North-Holland, 1989, pp. 61-128.
- Robbins, Lionel. *An essay on the nature and significance of economic science*. London: Macmillan, 1935.
- Robinson, Joan. *The economics of imperfect competition*. London: Macmillan, 1933.

The Institutions of Governance

By OLIVER E. WILLIAMSON*

The new institutional economics (NIE) is an idea whose time has come. That was evident to R. C. O. Matthews in 1986, who in his presidential address to the Royal Economic Society declared that the NIE was "one of the liveliest areas in our discipline" (1986 p. 903) and thereafter described it as a body of thinking based in two propositions: institutions matter, and institutions are *susceptible to analysis* (1986 p. 903). The proposition that institutions matter is embraced by institutional economists of all kinds, old and new. What distinguishes the NIE from earlier (and some contemporary) work on institutions is that institutions are susceptible to analysis. Older-style institutional economics was content to critique orthodoxy and collapsed for failure to advance a positive research agenda (Ronald Coase, 1984; Matthews, 1986). The NIE has responded to the challenge by (i) developing a comparative institutional logic of organization to which (ii) many applications and refutable implications accrue and in relation to which (iii) many empirical tests have been conducted and are broadly corroborative (Paul Joskow, 1991; Howard Shelanski and Peter Klein, 1995; Bruce Lyons, 1996; Keith Crocker and Scott Masten, 1996).

The institutions of principal interest to the NIE are the institutional environment (or rules of the game—the polity, judiciary, laws of contract and property [Douglas North, 1991]) and the institutions of governance (or play of the game—the use of markets, hybrids, firms, bureaus). Although it has been customary to work at one level at a time, each level informs the other, and recent work of a combined kind has appeared as applications to economic development and reform are attempted. Also, the study of economic institutions needs to make

provision for two background conditions: the condition of societal embeddedness (to which sociologists refer; [Mark Granovetter, 1985; Victor Nee, 1997]) and the attributes of human actors (Herbert Simon, 1985; Leda Cosmides and John Tooby, 1994, 1996). I return to these two background conditions in Section II. Suffice it to observe here that the NIE is, by its very nature, an interdisciplinary undertaking.

I. The Governance of Contractual Relations

Transaction-cost economics is located on the branch of the NIE that is predominantly concerned with governance, the branch that has its origins in Ronald Coase's treatment of firms and markets in his classic 1937 paper on "The Nature of the Firm." Rather than take the organization of economic activity in firms and markets as preexisting, defined largely by technology, Coase described firms and markets as alternative means for doing the very same thing. The allocation of activity as between markets and hierarchies was no longer taken as given, but needed to be derived. Should a firm make or buy? Which transactions go where and why? The firm was reconceptualized for these purposes as a governance structure (which is an organizational construction).

Much of the predictive content of transaction-cost economics works through the discriminating-alignment hypothesis: transactions, which differ in their attributes, are aligned with governance structures, which differ in their cost and competence, so as to effect a (mainly) transaction-cost economizing result. Implementing this requires that transactions, governance structures, and transaction-cost economizing all be described. What are the defining attributes of transactions? What are the attributes with respect to which governance structures differ? What main purposes are served by economic organization? How is transaction-cost economizing accomplished?

* Walter A. Hass School of Business, Department of Economics, Law School, University of California, Berkeley, CA 94720-1900.

According to John R. Commons (1932 p. 4), "the ultimate unit of activity ... must contain in itself the three principles of conflict, mutuality, and order. This unit is a transaction." Transaction-cost economics concurs that the transaction is the basic unit of analysis and regards governance as the means by which order is accomplished in a relation in which potential conflict threatens to undo or upset opportunities to realize mutual gains.

The problem of conflict upon which transaction-cost economics originally focused is that of bilateral dependency (Williamson, 1975; Benjamin Klein et al., 1978). Whereas the organization of transactions that are supported by generic investments is easy (classical market contracting works well because each party can go its own way with minimal cost to the other), potential problems arise when nonredeployable investments are made. Parties that are joined in a condition of bilateral dependency (by reason of asset specificity) and are confronted with contractual incompleteness (by reason of bounds on rationality) must confront strains (by reason of opportunism) when faced with the need to adapt cooperatively. How should such contracts be managed?

A comparative assessment of markets and hierarchies is needed. Taking adaptation to be the central problem of economic organization, of which autonomous and cooperative kinds are distinguished, markets enjoy the advantage in autonomous-adaptation respects, and the advantage shifts to hierarchies as the needs for cooperative adaptation build up. Discrete structural differences between markets and hierarchies in incentive intensity, administrative control, and contract-law regimes are responsible for these adaptive differences (Williamson, 1991).

Implicit in the foregoing and important to the transaction-cost economics enterprise is the assumption that contracts, albeit incomplete, are interpreted in a farsighted manner, according to which economic actors look ahead, perceive potential hazards, and embed transactions in governance structures that have hazard-mitigating purpose and effect. Also, most of the governance action works through private ordering, with courts being reserved for purposes of ultimate appeal. Furthermore,

although vertical integration is the archetype transaction out of which transaction-cost economics works, labor, capital, corporate governance, regulation/deregulation, vertical market restrictions, multinational and public-sector transactions are variations on a theme to which numerous public-policy ramifications accrue. Finally, although bilateral dependency (asset specificity) is a widespread condition, the basic transaction-cost setup, with its emphasis on *ex post* governance, applies to contractual hazards more generally.

Additional hazards that transaction-cost economics has begun to address include (i) the hazards of weak property rights (David J. Teece, 1986); (ii) measurement hazards of multiple-task (Bengt Holmstrom and Paul Milgrom, 1991), oversearching (Yoram Barzel, 1982; Roy Kenney and Klein, 1983), and multiple-principal (Avinash Dixit, 1996) kinds; (iii) intertemporal hazards, which can take the form of disequilibrium contracting, real-time responsiveness, long latency, and strategic abuse; (iv) hazards that accrue to weaknesses in the institutional environment (North and Barry Weingast, 1989; Brian Levy and Pablo Spiller, 1994; Weingast, 1995), which are pertinent to economic development and reform, and (v) weaknesses of probity, which are of special concern for what James Q. Wilson (1989) refers to as "sovereign transactions" (see also Williamson [1998]). To be sure, lenses other than transaction-cost economics can and have been brought to bear. It is nonetheless noteworthy that transaction-cost economics relates to all of these hazards in the following four respects: (i) all of these hazards would vanish but for bounds on rationality and opportunism, (ii) the magnitude of the hazards varies systematically with the attributes of transactions, (iii) *ex post* governance (as well as *ex ante* incentive alignment) is an important instrument in effecting hazard mitigation, and (iv) the discriminating-alignment hypothesis applies. Hazard mitigation through the *ex post* governance of incomplete contracts is the general rubric.

Whereas organizational variety was once ascribed principally to monopoly purpose and/or efficient risk-bearing, much of this variety is usefully interpreted in transaction-cost-economizing terms. It matters a lot, as it turns

out, whether firms are described as governance structures rather than production functions.

What came to be known as the "inhospitality tradition" in antitrust was the public-policy consequence of treating the firm as a production function (which is a technical construction). Because a technological justification for nonstandard and unfamiliar contractual and organizational practices was lacking under the neoclassical setup, such practices were presumed to be anticompetitive.

However, that is because the study of economics was needlessly truncated. Technological economies of scale or scope and related "technical or physical aspects" do not exhaust the possibilities. Organizational economies that are attributable to the alignment of governance structures with the attributes of transactions had been ignored or suppressed. Upon working through the logic of discriminating alignment, an altogether different rationale for vertical integration, vertical market restrictions, and other forms of nonstandard contracting took shape.

Many of the hitherto puzzling economic institutions of capitalism were interpreted more constructively as this broader approach to economizing progressed. Thus, although anticompetitive purpose remains a concern if the relevant market-power preconditions are satisfied, those are stringent preconditions. As a consequence, anticompetitive effect is better regarded as the exception rather than the rule.

The concepts of credible commitment and discriminating alignment that inform private-sector governance have also turned out to be instructive for understanding economic development and reform. Viewing the institutional arrangements (rules of the game) through the lens of contract and governance has helped, among other things, to disclose when and why privatization efforts will succeed or fail.

II. Background Conditions

The two background conditions to which I referred earlier are the conditions of embeddedness and the attributes of human actors. The first of these is antecedent to the polity and refers to societal features (norms, customs, mores, religion) which differ among

groups and nation states and operate as societal supports, or the lack thereof, for credible contracting. A recurring concern is when and why do reputation-effect mechanisms work well and poorly?

Turning to the attributes of human actors, Simon (1985 p. 303) takes the position that "Nothing is more fundamental in setting our research agenda and informing our research methods than our view of the nature of the human beings whose behavior we are studying." The huge literature on "psychology and economics," as recently summarized by Matthew Rabin (1996), speaks to many of the issues. Of special interest is research stemming from Daniel Kahneman and Amos Tversky (1982), in which the limits and biases of individual decision-makers in making probabilistic choices are revealed. Also pertinent are recent studies by cognitive anthropologists (Edwin Hutchins, 1995) and evolutionary psychologists (Cosmides and Tooby, 1996).

Hutchins's book on *Cognition in the Wild* makes the distinction between cognition "in the laboratory, where cognition is studied in captivity, and the everyday world, where human cognition adapts to its natural surroundings" (1995 p. xiv). Issues of organization are brought in by viewing individuals as "part of a larger computational system" (Hutchins, 1995 p. xv). Because, moreover, each generic form of organization has both strengths and weaknesses, organization needs to be studied as both problem and solution. Cognitive anthropology can help to inform both.

The evolutionary-psychology literature notes that cognitive methods "drawn from logic, mathematics, and probability theory" (Cosmides and Tooby, 1994 p. 319) and cognitive processes that had good survival properties for human agents operating in the Pleistocene sometimes differ. For example, framing problems of choice under uncertainty in frequentist terms, which is the manner in which they arise in the wild, rather than in point estimates, which is the way they are posed by statistical decision theory, is consequential. Cognitive biases and disabilities are often reduced when problems are presented in frequentist terms.

Evidence, moreover, that individuals on average are poor probabilists speaks only to the

mean. If there is variance in the aptitude to make probabilistic choices among members of the population, and if it is cost-effective to screen for differential competence, then an efficient alignment of differential competence to tasks can be effected through specialization (organization). Not only does organization matter in this respect, but it also relieves problems of limited cognitive ability by permitting composite tasks to be broken down into nearly separable parts, each of which is more tractable, thereafter to be recombined (Simon, 1962). The upshot is that the study of individual decision-making needs to be extended to make provision for the efficient deployment of individuals and groups within an organization. Cognitive anthropology, organization theory, and evolutionary psychology are all implicated.

III. Concluding Remarks

The papers and discussions at the inaugural conference of the International Society for New Institutional Economics in September 1997 record that the NIE continues to extend its analytical reach and inform new issues. New conceptual challenges arise as new applications are attempted. A deeper understanding of complex economic organization is in progress.

REFERENCES

- Barzel, Yoram. "Measurement Cost and the Organization of Markets." *Journal of Law and Economics*, April 1982, 25(1), pp. 27-48.
- Coase, Ronald H. "The Nature of the Firm." *Economica*, November 1937, 4(4), pp. 386-405.
- . "The New Institutional Economics." *Journal of Institutional and Theoretical Economics*, March 1984, 140(1), pp. 229-31.
- Commons, John R. "The Problem of Correlating Law, Economics, and Ethics." *Wisconsin Law Review*, December 1932, 8(1), pp. 3-26.
- Cosmides, Leda and Tooby, John. "Better than Rational: Evolutionary Psychology and the Invisible Hand." *American Economic Review*, May 1994 (*Papers and Proceedings*), 84(2), pp. 327-32.
- . "Are Humans Good Intuitive Statisticians After All?" *Cognition*, January 1996, 58(1), pp. 1-73.
- Crocker, Keith and Masten, Scott. "Regulation and Administered Contracts Revisited: Lessons from Transaction-Cost Economics for Public Utility Regulation." *Journal of Regulatory Economics*, January 1996, 9(1), pp. 5-39.
- Dixit, Avinash. *The making of economic policy: A transaction cost politics perspective*. Cambridge, MA: MIT Press, 1996.
- Granovetter, Mark. "Economic Action and Social Structure: The Problem of Embeddedness." *American Journal of Sociology*, November 1985, 91(3), pp. 481-501.
- Holmstrom, Bengt and Milgrom, Paul. "Multi-task Principal-Agent Analysis." *Journal of Law, Economics, and Organization*, Special Issue 1991, 7, pp. 24-52.
- Hutchins, Edwin. *Cognition in the wild*. Cambridge, MA: MIT Press, 1995.
- Joskow, Paul. "The Role of Transaction Cost Economics in Antitrust and Public Utility Regulatory Policies." *Journal of Law, Economics, and Organization*, Special Issue 1991, 7, pp. 53-83.
- Kahneman, Daniel and Tversky, Amos. "On the Study of Statistical Intuitions." *Cognition*, March 1982, 11(2), pp. 123-41.
- Kenney, Roy and Klein, Benjamin. "The Economics of Block Booking." *Journal of Law and Economics*, October 1983, 26(3), pp. 497-540.
- Klein, Benjamin; Crawford, T. A. and Alchian, A. A. "Vertical Integration, Appropriable Rents, and the Competitive Contracting Process." *Journal of Law and Economics*, October 1978, 21(2), pp. 297-326.
- Levy, Brian and Spiller, Pablo. "The Institutional Foundations of Regulatory Commitment: A Comparative Analysis of Telecommunications Regulation." *Journal of Law, Economics, and Organization*, October 1994, 10(2), pp. 201-46.
- Lyons, Bruce. "Empirical Relevance of Efficient Contract Theory: Inter-firm Contracts." *Oxford Review of Economic Policy*, Winter 1996, 12(4), pp. 27-52.

- Matthews, R. C. O. "The Economics of Institutions and the Sources of Growth." *Economic Journal*, December 1986, 96(384), pp. 903-18.
- Nee, Victor. "Sources of the New Institutionalism in Sociology." Unpublished manuscript, Cornell University, 1997.
- North, Douglass. "Institutions." *Journal of Economic Perspectives*, Winter 1991, 5(1), pp. 97-112.
- North, Douglass and Weingast, Barry. "Constitutions and Commitment: The Evolution of Institutions Governing Public Choice in 17th Century England." *Journal of Economic History*, December 1989, 49(4), pp. 803-32.
- Rabin, Matthew. "Psychology and Economics." Unpublished manuscript, University of California-Berkeley, 1996.
- Shelanski, Howard and Klein, Peter. "Empirical Research in Transaction Cost Economics: A Review and Assessment." *Journal of Law, Economics, and Organization*, October 1995, 11(2), pp. 335-61.
- Simon Herbert. "The Architecture of Complexity." *Proceedings of the American Philosophical Society*, December 1962, 106(6), pp. 467-82.
- . "Human Nature in Politics: The Dialogue of Psychology with Political Science." *American Political Science Review*, January 1985, 79(2), pp. 293-303.
- Teece, David J. "Profiting from Technological Innovation." *Research Policy*, December 1986, 15(6), pp. 285-305.
- Weingast, Barry R. "The Economic Role of Political Institutions: Market-Preserving Federalism and Economic Development." *Journal of Law, Economics, and Organization*, April 1995, 11(1), pp. 1-31.
- Williamson, Oliver E. *Markets and hierarchies: Analysis and antitrust implications*. New York: Free Press, 1975.
- . "Comparative Economic Organization: The Analysis of Discrete Structural Alternatives." *Administrative Science Quarterly*, June 1991, 36, pp. 269-96.
- . "Public and Private Bureaucracies." Unpublished manuscript, University of California-Berkeley, 1998.
- Wilson, James Q. *Bureaucracy*. New York: Basic Books, 1989.

Historical and Comparative Institutional Analysis

By AVNER GREIF*

Among the most fundamental questions of institutional economics are: Why do societies evolve along distinct institutional trajectories? Why do societies often fail to adopt the institutional structure of more successful ones? How may we examine the interrelations between the implicit and informal aspects of societies' institutions, on the one hand, and their explicit and formal aspects, on the other? A particular conceptual framework and empirical methodology, historical and comparative institutional analysis (HCIA), has recently been developed and employed to address these and other questions regarding the origins, nature, and implications of institutions and institutional change. What follows is a short elaboration on HCIA, its essence, and some preliminary insights.¹

HCIA is historical in its attempt to explore the role of history in institutional emergence, perpetuation, and change; it is comparative in its attempt to gain insights through comparative studies over time and space; and it is analytical in its explicit reliance on context-specific micro models for empirical analysis. HCIA conceptualizes institutions as the non-technologically determined *constraints* that influence social interactions and provide incentives to maintain regularities of behavior. It considers institutions that are *outcomes* emerging endogenously and that are *self-enforcing* in the sense that they do not rely on external enforcement. HCIA thus considers the *relevant*

rules of the game that actually constrain behavior in a society (as distinct from the technologically feasible rules) to be a self-enforcing outcome of forces, such as strategic interactions, evolutionary processes, and limits on cognition. These rules, in constituting part of a society's institutions, are complemented by self-enforcing constraints generated through interactions within these rules. An essence of HCIA is thus the examination of the factors determining the relevant rules of the game, the forces that make these rules self-enforcing, and the self-enforcing constraints on behavior that emerge within these rules. State-mandated rules, values, or social norms that actually constrain behavior, for example, are considered as outcomes rather than exogenous forces.

To advance such an examination, HCIA has so far mainly utilized the lens provided by the study of equilibria in a game-theoretical sense. The study of institutions through equilibrium analysis, the view of institutions as equilibrium constraints, enables examination of the static, endogenous, and self-enforcing constraints generated in strategic situations in the absence of external enforcement. Furthermore, it provides the basis for examining institutional origin and change as reflecting the interrelations among society's decision-makers, their past institutions, and the evolving environment within which they interact. The analysis, however, neither assumes the appropriateness of using standard game-theoretic solutions nor the prevalence of an institution with particular attributes, such as efficiency or equity. Rather, at the heart of HCIA's research strategy is an inductive, empirical analysis regarding the relevance of particular institutions based on evaluating and synthesizing micro-level historical and comparative evidence and insights from context-specific, micro theoretical models. The sensitivity of outcomes to specifications and the indeterminacy of equilibrium indicate the importance of integrating historical and comparative studies in pursuing empirical institutional analysis.

* Department of Economics, Stanford University. I benefited from the insightful comments by Masahiko Aoki and Paul Milgrom.

¹ This approach builds on existing lines of research in economics, the new institutional economics, sociology, and political science. For previous elaborations see Masahiko Aoki (1996), particularly on the use of evolutionary game theory to examine complementary organizations through comparative institutional analysis, and Greif (1997), particularly on the use of game theory for comparative studies of organizations and beliefs through historical institutional analysis (also see Aoki [1998] and Greif [1998]).

There are two lines of analysis within HCIA. The first considers the impact of the internalization of traits through evolutionary process and learning on the set of the relevant rules. It utilizes evolutionary game theory and learning models to study the process through which decision-makers with particular traits, such as specific organizational features, preferences, or habits, emerge and the constraints on behavior that their interactions entail. It further examines the complementarities among traits in various spheres of economic activities, and between them and government regulations and rules. The focus of the empirical studies in this line of analysis has been mainly on situations in which traits are relatively easy to observe, such as the emergence of firms with particular capabilities and their institutional complementarity with financial systems, employment relations, and government regulations (Aoki, 1994, 1995; Tetsuji Okazaki and Masahiro Okuno-Fujiwara, 1996). Related, more theoretical works also have examined such issues as the emergence of conventions, customary property rights allocations, and preference traits (e.g., Robert H. Frank, 1987; Robert Sugden, 1989; H. P. Young, 1993).

The second line of analysis considers the impact of strategic interactions and exogenous and endogenous cultural features, beliefs, social structures, and cognition (such as awareness) on the set of the relevant rules. It employs mainly (classical) game theory, particularly the theory of repeated games, and concentrates on the origin and implications of (nontechnologically determined) "organizations," and the constraints implied by beliefs prominent in a society regarding behavior on and off the path of play. Organizations alter the set of the relevant rules of the game by constituting a new player (the organization itself), changing the information available to the players, or changing payoffs associated with certain actions. Examples of such organizations include the merchant guild, the firm, the bank, and the credit bureau. While these organizations alter the set of the relevant rules of the game in the societies in which they prevail, and some organizations are strategic players, their emergence nevertheless represents actions taken, in the appropriate meta-game by those who established them.

Conceptually, these actions differ from other actions only by potentially having a "profound" impact, implying a qualitative change in the set of possible institutional constraints relative to those possible in the same game in the absence of the organization under consideration.

The empirical methodology employed in this line of analysis reflects its concentration on the identification of relevant organizations and beliefs on and off the path of play and its attempt to evaluate, rather than assume, the relevance of game theory. The point of departure for the analysis is the identification of relevant institutions, sets of self-enforcing expectations and organizations and related features, such as behavior and social structures, relevant in the particular historical episode under consideration. It does not begin by contemplating the set of theoretically feasible institutions and choosing among them based on some deductive theory or objective criteria, because the extent of knowledge, rationality, and cognition is to be evaluated rather than assumed.

An hypothesis regarding the relevance of a particular institution is formulated based on a micro-level, detailed examination of the evidence. It is expressed with the assistance of a context-specific model whose details are based on the evidence and whose robustness is evaluated, particularly regarding aspects that are not well reflected empirically. Furthermore, since game-theoretical formulation is the benchmark of empirical analysis rather than the mold, the hypothesis regarding the relevance of a particular institution and its game-theoretical formulation has to be empirically substantiated. Substantiation is particularly important because game theory provides a limited guide to equilibrium selection, and it entails contrasting predictions implied by the theoretical analysis with the historical and comparative observations and data.

Following substantiation, the factors leading to the *emergence* of the institution, as well as its *implications*, are examined. This examination rests upon the game-theoretical insight that multiple equilibria are likely to exist in a given strategic situation. It indicates that the study of institutional emergence can benefit from considering their interrelations with

noneconomic factors: the historical, cultural, social, and political aspects of the particular society under consideration. Similarly, recognizing that historical actors have a limited rationality implies that studying institutional change and innovations requires going beyond the conceptual confines of game theory. HICA postulates that institutional dynamics might not be an optimal response to the changing environment, a reflection of random mutation in organizations or beliefs, or a change in the power of a particular political actor. Rather, it explores the possibility that institutional changes also reflect the limits on rationality, cognition, and knowledge, and the incentive for institutional innovations, adoption, and change implied by the existing institutions and circumstances.

The empirical studies in this second line of analysis have focused on many topics, such as the formal and informal institutional foundations of the market and informal systems for contract enforcement (Greif, 1989, 1993, 1997a; Marcel Fafchamps, 1996; Karen Clay, 1997); the role of culture in the emergence and perpetuation of distinct institutional and organizational trajectories (Greif, 1994b); the institutional foundation of the state and the political foundations of market economies (Douglass C. North and Barry R. Weingast, 1989; Greif, 1997b; Weingast, 1997); the interrelations among social structures, culture, and economic and political institutions (Greif, 1994b, 1997b); and the emergence, perpetuation, and change of alternative formal and informal financial systems and distinct institutions governing labor relations (Timothy W. Guinnane, 1994; Aoki and Serdar Dinç, 1997; Chiaki Moriguchi, 1997). The focus of the more theoretical works has been on issues such as the role of various organizations in facilitating cooperation, organizational complementarities, and the institutional foundations of the state (e.g., Paul Milgrom et al., 1990; Milgrom and John Roberts, 1992; Randall L. Calvert, 1996; Robert Gibbons and Andrew Rutten, 1997).

Studies in HICA highlight the nature, origin, and implications of institutions and institutional change in particular historical episodes. Yet some insights have emerged in more than one work, suggesting that they may be general in

nature. These insights indicate, for example, that complementarities among past economic institutions impact institutional evolution (Aoki, 1995; Greif, 1994b; Moriguchi, 1997). By providing information, enabling the conditioning of strategies on social identity, and making social sanctions feasible, initial social structures permit the emergence of particular self-enforcing economic and political institutions whose functioning further influence these social structures (e.g., Greif, 1989, 1994b, 1997a, b). Cultural beliefs embedded in existing institutions direct the process of organizational innovation and adoption, as well as cultural and social evolution, while intentional and unintentional organizational learning and innovations are shaped by the incentives provided by current institutions and organizational failure (e.g., Greif, 1994b, 1997b; Greif et al., 1994; Guinnane, 1994).

Together, these insights indicate that a society's institutions are a complex in which informal, implicit institutional features interrelate with formal, explicit features in creating a coherent whole. These interrelations direct institutional change and cause this institutional complex to resist change more than its constituting parts would have done in isolation. Hence, this institutional complex is not a static optimal response to economic needs. Rather, it is a reflection of an historical process in which past economic, political, social, and cultural features interrelate and have a lasting impact on the nature and economic implications of a society's institutions (e.g., Greif, 1994a, b, 1997; Aoki, 1995; Moriguchi, 1997).

HICA thus reveals both the forces that lead societies to evolve along distinct institutional trajectories and the sources of the difficulties that societies face in adopting the institutions of more successful ones. More broadly, it indicates the importance of examining a society's self-enforcing endogenous institutions as products of an historical process in which past institutional, economic, political, social, and cultural features interact in shaping the nature of contemporary institutions and their evolution. Furthermore, HICA's achievements prove the feasibility of conducting such an analysis based on integrating game-theoretical and empirical, historical, and comparative studies.

REFERENCES

- Aoki, Masahiko. "The Contingent Governance of Teams: Analysis of Institutional Complementarity." *International Economic Review*, August 1994, 35(3), pp. 657-76.
- . "Organizational Conventions and the Gains from Diversity: An Evolutionary Game Approach." Working paper, Stanford University, 1995.
- . "Towards a Comparative Institutional Analysis: Motivations and Some Tentative General Insights." *Japanese Economic Review*, March 1996, 47(1), pp. 1-19.
- . *Toward a comparative institutional analysis*. Cambridge, MA: MIT Press, 1998 (forthcoming).
- Aoki, Masahiko and Dinç, Serdar. "Relational Financing as an Institution and Its Viability under Competition." Working paper, Stanford University, 1997.
- Calvert, Randall L. "Rational Actors, Equilibrium, and Social Institutions," in Jack Knight and Itai Sened, eds., *Explaining social institutions*. Ann Arbor: University of Michigan Press, 1996, pp. 57-94.
- Clay, Karen. "Trade Without Law: Private-Order Institutions in Mexican California." *Journal of Law, Economics, and Organization*, April 1997, 13(1), pp. 202-31.
- Fafchamps, Marcel. "Market Emergence, Trust and Reputation." Working paper, Stanford University, 1996.
- Frank, Robert H. "If Homo Economicus Could Choose His Own Utility Function, Would He Want One with a Conscience?" *American Economic Review*, September 1987, 77(4), pp. 593-604.
- Gibbons, Robert and Rutten, Andrew. "Hierarchical Dilemmas: Social Order with Self-Interested Rulers." Working paper, Cornell University, 1997.
- Greif, Avner. "Reputation and Coalitions in Medieval Trade: Evidence on the Maghribi Traders." *Journal of Economic History*, December 1989, 49(4), pp. 857-82.
- . "Contract Enforceability and Economic Institutions in Early Trade: The Maghribi Traders' Coalition." *American Economic Review*, June 1993, 83(3), pp. 525-48.
- . "Trading Institutions and the Commercial Revolution in Medieval Europe," in Abel Aganbegyan, Oleg Bogomolov, and Michael Kaser, eds., *Economics in a changing world*, Vol. 1. London: Macmillan, 1994a, pp. 115-25.
- . "Cultural Beliefs and the Organization of Society: A Historical and Theoretical Reflection on Collectivist and Individualist Societies." *Journal of Political Economy*, October 1994b, 102(5), pp. 912-50.
- . "Microtheory and Recent Developments in the Study of Economic Institutions Through Economic History," in David M. Kreps and Kenneth F. Wallis, eds., *Advances in economic theory*, Vol. II. Cambridge: Cambridge University Press, 1997, pp. 79-113.
- . "On the Historical Development and Social Foundations of Institutions that Facilitate Impersonal Exchange." Working paper, Stanford University, 1997a.
- . "Self-Enforcing Political System and Economic Growth: Late Medieval Genoa." Working paper, Stanford University, 1997b.
- . *Genoa and the Maghribi traders: Historical and comparative institutional analysis*. Cambridge: Cambridge University Press, 1998 (forthcoming).
- Greif, Avner; Milgrom, Paul and Weingast, Barry. "Coordination, Commitment, and Enforcement: The Case of the Merchant Guild." *Journal of Political Economy*, August 1994, 102(4), pp. 912-50.
- Guinnane, Timothy W. "A Failed Institutional Transplant: Raiffeisen's Credit Cooperatives in Ireland, 1894-1914." *Explorations in Economic History*, January 1994, 31(1), pp. 38-61.
- Milgrom, Paul R.; North, Douglass and Weingast, Barry R. "The Role of Institutions in the Revival of Trade: The Medieval Law Merchant, Private Judges, and the Champagne Fairs." *Economics and Politics*, March 1990, 2(1), pp. 1-23.
- Milgrom, Paul R. and Roberts, John. *Economics, organization, and management*. Englewood Cliffs, NJ: Prentice Hall, 1992.
- Moriguchi, Chiaki. "Evolution of Employment Systems in the US and Japan: 1900-60. A Comparative Historical Analysis."

- Unpublished manuscript, Stanford University, 1997.
- North, Douglass C. and Weingast, Barry R. "Constitutions and Commitment: Evolution of Institutions Governing Public Choice," *Journal of Economic History*, December 1989, 49(4), pp. 803–32.
- Okazaki, Tetsuji and Okuno-Fujiwara, Masahiro. "Evolution of Economic Systems: The Case of Japan," in Masahiko Aoki and Huijiro Hayami, eds., *The institutional foundations of economic development in East Asia*. London: Macmillan, 1996.
- Sugden, Robert. "Spontaneous Order." *Journal of Economic Perspectives*, Fall 1989, 3(4), pp. 85–97.
- Weingast, Barry R. "The Political Foundations on Democracy and the Rule of Law." *American Political Science Review*, June 1997, 91(2), pp. 245–63.
- Young, H. P. "The Evolution of Conventions." *Econometrica*, January 1993, 61(1), pp. 57–84.

Norms and Networks in Economic and Organizational Performance

By VICTOR NEE*

Analyzing institutions and the part they play in governing economic action has comprised the central problem addressed by the new institutional economics. This paradigm has made impressive progress in explaining a wide array of economic activity by extending the tools of standard economic theory to examine institutions. Thus far, new institutional economics has emphasized understanding the part played by formal institutional arrangements: contracts, property rights, laws, regulations, and the state. In recent years, new institutionalists have also incorporated a focus on informal constraints, principally norms and networks. These studies examine how informal constraints provide a framework for collective action and furnish an alternative mechanism in enforcing the rules of the game and facilitating transactions between economic actors (e.g., Robert Ellickson, 1991). But they do not examine the manner in which informal and formal constraints *combine* to shape the performance of organizations and economies.

The economics of organizations pioneered by Ronald Coase (1937) and Oliver Williamson (1975) proffers predictive hypotheses as to the forms of organization and governance structure likely to emerge. On a larger canvass, Douglass North (1981) extended his theory of the state to explain the political conditions that give rise to efficient property rights favorable to economic growth. These theories emphasize the causal significance of formal constraints in making predictions about economic and organizational performance.

However, explanations by reference to institutional effects are problematic without a theory of the nature of the relationship between formal and informal constraints. What is the nature of the relationship between infor-

mal and formal constraints in explaining variation in performance? Informal constraints embedded in norms and networks, operating in the shadows of formal organizational rules, can both limit and facilitate economic action. Informal constraints can give rise to inefficient allocation of resources when private entrepreneurial networks collude to secure resources from government for their group, resulting in structural rigidities and economic stagnation (Mancur Olson, 1982). They can also facilitate the growth of a new industry by providing a framework for trust and collective action, as documented in a case study of Thomas Edison and the rise of the electric utility industry (Patrick McGuire et al., 1993).

Unless the nature of the relationship between informal and formal constraints is better specified, the inclusion of informal constraints by new institutionalists contributes to a problem of indeterminacy. Transaction-cost reasoning based on analysis of formal constraints may identify the optimal institutional arrangement from the vantage point of the corporate actor, but this may be abraded by informal norms and networks operating to realize interests and preferences of individuals in subgroups. Uncertainty about whether informal constraints in a particular institutional environment will function in opposition or in harmony with the formal rules accentuates a difficulty of prediction in the new institutional economics. How is it possible to extend the same transaction-cost reasoning to the domain of ongoing social relationships, where economic theory runs into predictable trouble in accounting for something as elusive as sentiment and identities?

I. Informal Norms and Networks: Examples

In transition economics, rulers implemented far-reaching transformation of formal rules of the game to institute market economies.

* Department of Sociology, 312 Uris Hall, Cornell University, Ithaca, NY 14853.

Economic analyses of transition economies tend to focus on the changes in the institutional framework resulting in (and from) the dismantling of state control over economic activity—for good reason, as those changes shift incentives away from bargaining with bureaucrats in political markets toward seeking profit and gain in economic markets. But whether rulers followed blueprints of capitalist transition provided by Western economists or pursued a trial-and-error evolutionary approach, the behavior of economic actors frequently bears little resemblance to the legitimate courses of action stipulated by the formal rules. Instead, networks based on personal connections serve to organize market-oriented economic behavior according to informal norms reflecting the private expectations of entrepreneurs and politicians. They act in the shadow of the state, often at odds with the goals formulated by rulers. In China, informal privatization and local arrangements have contributed to a remarkable two decades of sustained economic growth. But in Russia, mafia-like business networks have operated to obstruct Boris Yeltsin's efforts at building a modern market economy. What accounts for the difference in results? One thing is certain: a one-sided focus on either formal or informal constraints would miss the boat in understanding economic behavior in transition economies.

In advanced market economies, informal constraints influence productivity to a larger extent than has been readily acknowledged by economists. The workaday norms and network structure of workers have a decisive effect on organizational performance, as Fritz Roethlisberger and William Dickson's study (1939) of a work group in the Western Electric Company's Hawthorne Plant documents. This work group's productivity was not regulated by management, since the foreman assigned to the group chose not to enforce formal organizational rules so as not to sacrifice his good relationships with the men in the work group. Instead, he chose to side with workers and to "wink" at activities in the workshop that violated formal rules. In practice, group performance was regulated by an informal output norm, which limited production to two pieces of equipment a day. This output was considered satisfactory by management; however, it was not as great as it could have been

if worker fatigue were the limiting factor. Workers enforced the "restriction of output" by submitting offenders to merciless ridicule. They called a fellow worker who exceeded the informal output norm a "rate-buster" or "speed king," while they labeled someone who fell below the norm a "chiseler," for cutting down the earnings of the group. "Binging" (hitting as hard as possible on the upper arm) was meted out to fellow workers who worked either too fast or too slow. Workers who consistently conformed to the informal output norm enjoyed higher informal rank, reflected in social approval from fellow workers and a position of centrality in the network structure, while those who regularly violated the output norm were ostracized.

Although these examples (work group and transition economies) are drawn from very different institutional environments, there is a common thread. In both kinds of contexts, the formal constraints have established the parameters of legitimate action, providing the institutional mold within which emergent norms and networks operate. This is seen in transition economies where the state still controls resources needed by entrepreneurs who must manage their firms in quasi-markets. Not surprisingly partial reform and quasi-markets result in "crony capitalism," characterized by strategic network ties between entrepreneurs and politicians. Similarly, in the Hawthorne Plant of the Western Electric Company, the informal norm emerged in response to the group piece-rate program designed by management to encourage workers to work up to the limit imposed by fatigue. Workers instead enforced the informal norm as a "satisficing" strategy in opposition to the formal rules. In both kinds of contexts, the informal overwhelmed the formal rules of the game in shaping economic action. Network ties provided the basis for trust and identity in close-knit groups, and informal norms enabled actors to engage in collective action to realize their preferences and interests.

II. The Relationship between Formal and Informal Constraints

Herein lies the problem. Formal rules are produced and enforced by organizations such

as the state and firm to solve problems of collective action through third-party sanctions, while informal norms arise out of networks and are reinforced by means of ongoing social relationships. To the extent that members of networks have interests and preferences independent of what rulers and entrepreneurs want, the respective contents of informal norms and formal organizational rules are likely to reflect opposing aims and values.

Unlike formal rules, the monitoring of informal norms is intrinsic to the social relationship, and enforcement occurs informally as a by-product of social interaction. Norms are implicit or explicit rules of expected behavior that embody the interests and preferences of members of a close-knit group or a community. The norm of "publish or perish," for example, articulates the preference of senior faculty and administrators in research universities where the ranking of a department and university is based on the research productivity of its faculty. The formal procedures of promotion review, providing an institutionalized mechanism to enforce a high standard of faculty quality, preceded the emergence of the "publish or perish" norm. The formal rules emphasize teaching and research. But according to the norm of "publish or perish," individual faculty members are expected to devote their best effort to research and writing. Although teaching and department citizenship are valued, the norm makes the priority clear. Faculties strive to maintain a high standard of research productivity because it is in their collective interest to secure a high national ranking for their departments. But the everyday interactions of faculty members enforce the informal norm of "publish or perish" to the extent that those who are "active" enjoy social approval and higher social rank, and those who are unproductive are marginalized. Insofar as norms help solve the problem of coordination and collective action, they enable actors to capture the gains from cooperation, which in the case of "publish or perish" is a higher national ranking, research funding, and the capacity to attract better students.

The proposition that individuals jointly produce and uphold norms to capture the gains of cooperation opens the way to specifying the relationship between informal and formal con-

straints (Nee and Paul Ingram, 1998). It is consonant with Ellickson's (1991) welfare-maximizing hypothesis for workaday norms: "Members of a close-knit group develop and maintain norms whose content serves to maximize the aggregate welfare that members obtain in their workaday affairs with one another." Both assume that norms are ideas that arise from the problem-solving activities of human beings in their strivings to improve their chances for success (the attainment of rewards) through cooperation. Informal norms arise through trial and error and are adopted by members of a group when they result in success. Whether members of a group are individually rewarded governs the selection of a norm. The core assumption behind each proposition is that informal norms embody interests and preferences that can only be realized by means of collective action.

III. Congruent, Decoupled, and Opposition Norms

When the formal rules of an organization are perceived to be congruent with the preferences and interests of actors in subgroups, the relationship between formal and informal norms will be closely coupled. The close coupling of informal norms and formal rules is what promotes high performance in organizations and economies. When the informal and formal rules of the game are closely coupled, they are mutually reinforcing. This is illustrated in the case of research universities in the close coupling between formal review procedures gauging and rewarding research productivity and the informal norm of "publish or perish." It is also seen in the congruence between informal norms of fair play and formal rules in competitive games. When informal and formal norms are closely coupled, it is often difficult to demarcate the boundaries between informal and formal social control. Close coupling of informal and formal constraints results in lower transaction costs because monitoring and enforcement can be accomplished informally. The cost of social rewards to achieve conformity to norms is low because it is produced spontaneously in the course of social interactions in networks of personal relations. By contrast, the greater the

reliance on formal sanctions, the higher the transaction costs involved in maintaining compliance.

When the formal rules are at variance with the preferences and interests of subgroups in an organization, a decoupling of the informal norms and practical activities, on the one hand, and the formal rules, on the other hand, will occur. As John Meyer and Brian Rowan (1977) observe, decoupling "enables organizations to maintain standardized, legitimating, formal structures, while their activities vary in response to practical considerations." For certain types of organizations, particularly those (such as public schools and government agencies) for whose output there is not a competitive market, formal organizational rules will be largely ceremonial, designed to satisfy external constituents that provide the organization with legitimacy. Independent of this ceremonial formal structure, informal norms arise to guide the day-to-day business of the organization.

Informal norms evolve into "opposition norms" if institutions and organizational sanctions are weak relative to contradicting group interests. Opposition norms encourage individuals to directly resist formal rules. In state socialist societies where the state-managed economy was widely perceived to be at odds with the interests of economic actors, opposition norms emerged to organize what came to be known as the illegal "second economy." Elsewhere too, when the organizational leadership and formal norms are perceived to be at odds with the interests and preferences of actors in subgroups, informal norms opposing formal rules will emerge to "bend the bars of the iron cage" of the formal organizational rules. Opposition norms have the most negative implication for performance. They give rise to organizational conflict and factionalism and often result in low morale.

IV. Concluding Remarks

Institutional design requires a combination of poetry and science. The cold rationalist view based on the extension of standard economic theory to analyze the workings of institutions is effective so far as the formal organizational rules are concerned. However, in the domain of informal norms and networks of ongoing social

relationships, a poet's insight into the human condition may prove to be as useful in institutional design as science. In a recent Gallup Organization survey of 55,000 workers to match employee attitude to organizational performance, four attitudes were found to be strongly correlated with company results: that workers feel they are provided the opportunity to do their best work, that they feel their opinions matter, that they are confident that fellow workers are also committed to quality, and that they sense their effort contributes directly to the company's success (Linda Grant, 1998). All of these matters involve sentiment and identity, which confer a sense of self-worth and purpose realized in the workplace. In other words, informal norms are not always utilitarian in content as they also embody more intangible states of sentiment and identity that arise from ongoing social relationships.

Notwithstanding, to the extent that the formal rules are consonant with the preferences and interests of organizational actors, informal processes of social control largely subsume the cost of monitoring and enforcement. It is this circumstance that affords for lower transaction costs, often leading to high economic and organizational performance. When rulers and entrepreneurs ignore or ride roughshod over this principle, they are likely to confront opposition norms and networks that organize resistance to their goals.

REFERENCES

- Coase, Ronald H. "On the Nature of the Firm." *Economica*, November 1937, 4, pp. 386-405.
- Ellickson, Robert. *Order without law*. Cambridge, MA: Harvard University Press, 1991.
- Grant, Linda. "Happy Workers, High Returns." *Fortune*, 12 January 1998, 137(1), pp. 81-92.
- McGuire, Patrick; Granovetter, Mark and Schwartz, Michael. "Thomas Edison and the Social Construction of the Early Electricity Industry in America," in Richard Swedberg, ed., *Explorations in economic sociology*. New York: Russell Sage Foundation, 1993, pp. 213-48.
- Meyer, John W. and Rowan, Brian. "Institutionalized Organizations: Formal Structure as

- Myth and Ceremony. *American Journal of Sociology*, September 1977, 83(2), pp. 340-63.
- Nee, Victor and Ingram, Paul. "Embeddedness and Beyond: Institutions, Exchange and Social Structure," in M. Brinton and V. Nee, eds., *The new institutionalism in sociology*. New York: Russell Sage Foundation, 1998, pp. 19-45.
- North, Douglass C. *Structure and change in economic history*. New York: Norton, 1981.
- Olson, Mancur. *The rise and decline of nations: Economic growth, stagflation, and social rigidities*. New Haven, CT: Yale University Press, 1982.
- Roethlisberger, Fritz J. and Dickson, William J. *Management and the worker*. Cambridge, MA: Harvard University Press, 1939.
- Williamson, Oliver E. *Markets and hierarchies: Analysis and antitrust implications*. New York: Free Press, 1975.

WHAT WE GET FOR HEALTH-CARE SPENDING[†]

Technological Change in Heart-Disease Treatment: Does High Tech Mean Low Value?

By MARK MCCLELLAN AND HARUKO NOGUCHI*

Trends in death and disability from heart disease in the United States over the past 30 years have been remarkable, accounting for the bulk of the substantial increase in longevity of older Americans. The elderly have the highest mortality rate from heart disease and other chronic illnesses and have also experienced the greatest absolute reduction in heart-disease death rates. Figure 1 shows that these improvements are reflected in the trends in outcomes of elderly patients hospitalized with heart attacks, a principal cause of death for patients with coronary heart disease. Between 1984 and 1994, one-year mortality for heart-attack patients declined by almost 10 percentage points (approximately one-fourth). Among survivors, the rate of rehospitalization with complications of heart disease, including heart failure and recurrent symptoms of coronary artery disease, did not rise substantially, suggesting that the improved survival was not in poor health status.

Figure 2 shows that this improvement has come at the cost of additional social resources. Real Medicare expenditures per patient for hospital care in the year following the attack have increased by more than 4 percent per year, with more rapid growth in recent years. These outcome and expenditure trends illus-

trate a central theme in the health economics of developed countries: improvements in the health of older populations have been accompanied by substantial growth in medical expenditures. For the case of heart attacks, the increase in expenditures appears to be attributable to changes in the nature and quantity of medical services provided (i.e., technological change) and not to real growth in the prices of medical services (see David Cutler and McClellan, 1998). Cutler et al. (1996), using reasonable valuations of the health improvements accompanying this technological change, concluded that the expenditure growth from 1984 to 1991 was worthwhile.

Even if technological change has been welfare-increasing on average, many economists have questioned whether changes in medical treatment are occurring in a way that improves the productivity of the medical system as much as they might. For example, McClellan et al. (1994) found that the marginal value of some intensive cardiac procedures is close to zero. Understanding whether technological change is occurring efficiently in health care requires a closer examination of how it occurs. But the study of health-care production remains largely a black box for most economists, demographers, and epidemiologists who have chronicled the substantial improvements in health and accompanying growth in medical expenditures. The black-box approach to changes in health-care production seems to represent something of a missed opportunity; because an enormous research literature in the biomedical sciences and in health-care cost-effectiveness analysis has examined the effects of particular medical technologies on changes in outcomes and, increasingly, medical expenditures. With few exceptions, these research literatures have not been linked.

[†] *Discussants:* David Meltzer, University of Chicago; Alan M. Garber, Stanford University.

* McClellan: Department of Economics and Department of Medicine, Stanford University, MC 6072, Stanford, CA 94305-6072, and National Bureau of Economic Research; Noguchi: National Bureau of Economic Research, Stanford, CA 94305. We thank Alan Garber and David Meltzer for helpful comments and the National Institute on Aging and the Health Care Financing Administration for research support. Data on Medicare beneficiaries are used by permission of the Health Care Financing Administration.

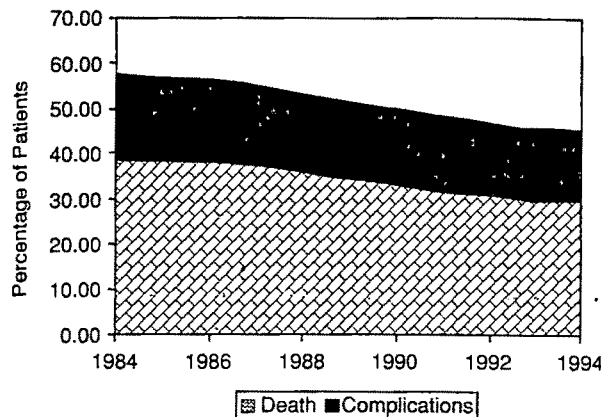


FIGURE 1. ONE-YEAR OUTCOMES AFTER HEART ATTACK, U.S. ELDERLY POPULATION

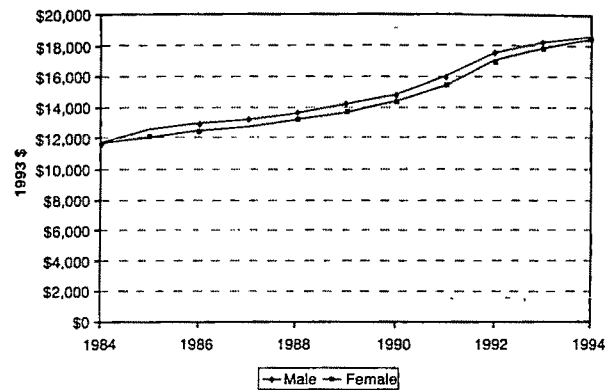


FIGURE 2. ONE-YEAR HOSPITAL EXPENDITURES AFTER HEART ATTACK, U.S. ELDERLY POPULATION

A principal reason is the enormous number of drugs, devices, and medical procedures that may comprise “technological change” in the production of care for a health problem. Reliable, simple measures of capital, labor, and materials inputs in production are hard to come by in health care, in large part as a result of the complexity and rapidity of technological change. The alternative approach, more careful analysis of the particular types of technological change, appears to have been viewed as too “field-specific” by most economists. Exceptions include descriptive studies of technological change by Annetine Gelijns and Nathan Rosenberg (1994). Many randomized and nonrandomized studies, as well as meta-analyses and synthetic reviews, of the effects of individual technologies, have been published in the medical literature. But virtually no investigators have examined the consequences of different kinds of technological change for population outcomes or resource use.

I. Types of Technological Change in Health Care

To provide a framework for assessing technological change, we develop a two-part typology. We divide innovations into *low-tech* and *high-tech* treatments, and we distinguish technological change that consists of *new* innovations from technological change that consists of changes in use or *diffusion* of existing technologies.

We define high-tech innovations as those with large fixed or marginal costs when they

are applied. (Far more medical technologies have large fixed development costs.) High-tech interventions for heart-attack patients include many that have received prominent press coverage, including cardiac catheterization to image the blood flow in the heart, coronary-artery bypass grafting to restore blood flow to the heart via intensive open-heart surgery, Swan-Ganz catheters to monitor heart function quantitatively in an intensive-care-unit setting, and balloon pumps to assist the function of the heart. Angioplasty, a technology that became widely used in the 1980’s, involves using a different type of balloon at the end of a catheter to open up narrowed heart blood vessels. In the 1990’s, this technology became more widely used in the immediate (same-day) treatment of a heart attack, to remove the blockage actually causing the attack. All of these technologies involve substantial high-skilled labor input, including specialized cardiologists, cardiac surgeons, cardiac nurses, and procedure technicians. They generally must be used in carefully controlled, dedicated settings. As a result, maintaining the capacity to provide these technologies is costly.

We define low-tech innovations as those with relatively low fixed and marginal costs, so that in principle they could be provided by virtually any medical facility. Though “low-tech” innovations have received less popular coverage, they have also been a major component of technological change. For example, clinical trials in the 1980’s showed that giving

aspirin during a heart attack leads to a substantial reduction in acute mortality from the attack (H. D. Lewis et al., 1983). Other innovations in acute drug treatments include the use of thrombolytic drugs, also known as clot-busters, which have also been shown in clinical trials to reduce mortality by dissolving the blockage causing an attack (ISIS-2 Collaborative Group, 1988). In addition, changes in low-tech interventions by medical providers, including patient monitoring, counseling and "caring" activities, and rehabilitation both in and out of a hospital, may also improve outcomes.

The low- and high-tech innovations that are used widely enough to contribute substantially to outcome or expenditure trends generally are not "new." Far more drugs, devices, and procedures are developed or proposed than become widely used in medical practice, and widespread use generally requires some time to occur. Thus, technology diffusion, not new technology development, appears to be the major proximate cause of expenditure and outcome effects. For example, after publication of the trial results on aspirin and thrombolytic drugs, their use in heart-attack treatment increased gradually over a period of many years, and even now many patients who may benefit from these drugs are not receiving them (Harlan Krumholz et al., 1997). The high-tech heart procedures that were associated with most of the growth in Medicare expenditures for heart-attack patients in the 1980's (cardiac catheterization, bypass surgery, and angioplasty) were all developed in the 1970's or earlier.

II. Determining the Importance of Different Types of Technological Change

Has the value of high- and low-tech innovations differed? This question has implications for understanding medical productivity and designing policies to increase it. For example, if high-tech innovations tend to be of low value, then regulatory and economic incentives that reduce their diffusion, such as "certification of need" regulation in New York and other states, and strict capital budgeting in other countries, may be desirable. Is technology diffusion too rapid or too slow? If

diffusion is important and occurs too slowly or too quickly because of inadequate knowledge about the advantages or limitations of technologies, then policies to influence experience and promote knowledge about their appropriate use may be critical. For example, the publication of medical practice guidelines and the evaluation of medical providers on the basis of such guidelines may have important consequences for outcomes and expenditures.

No randomized controlled trial could be performed to evaluate the impact of all of the many treatments that make up these different kinds of technological change. An alternative approach involves simulating how much of the observed improvements in outcomes and expenditures are plausibly explained by each type. Using evidence on changes in patient risk factors, Maria Hunink et al. (1997) concluded that 43 percent of the decline in coronary-heart-disease mortality between 1980 and 1990 resulted from improvements in acute treatment, including both low-tech and high-tech innovations. Another 29 percent of the improvement was attributed to "secondary" prevention of mortality, through the reduction of risk factors in patients with known heart disease. Much of this improvement probably reflected low-tech innovations such as greater use of medications to reduce blood pressure and cholesterol levels. In a broader overview of the changes in mortality from heart disease in the 20th century, Eugene Braunwald (1997) concluded that both low-tech innovations and high-tech innovations probably contributed significantly to the outcome improvements of the 1980's and 1990's. However, these studies did not try to quantify the effects of particular medical technologies or types of technologies.

Paul Heidenreich and McClellan (1998) combined a review of metaanalyses of treatment effects with a review of changes in the treatment of heart attacks to provide further evidence on the relative contributions of different types of technological change. They found that low-tech innovations probably accounted for the bulk of the heart-attack mortality decline between 1975 and 1995. But the limited published data on medical practices and the effects of treatments in differ-

ent types of patients, and in combination with other technologies, make the results of this study speculative as well. For example, few firm conclusions about contributions to long-term outcomes are possible, and there is relatively little published data on the effects of medical treatments on overall medical expenditures.

Another source of evidence is actual cost and outcome consequences of changes in medical practice. Because patients who receive different treatments probably differ in unmeasured dimensions, such comparisons require care to avoid selection bias. For example, using data on the U.S. elderly population with heart attacks, we have studied whether hospitals that differ in their access to new technologies have any consequences for patient welfare.

To illustrate the empirical method, consider a two-period model with innovation occurring between periods. We distinguish four types of technological innovations: low-tech, without knowledge or experience barriers; low-tech, with barriers; high-tech; and new technology of any type. If innovation is low-tech, with no systematic knowledge or diffusion barriers, then outcome improvements should take place at all hospitals. For example, if all hospitals become more likely to use aspirin in the acute treatment of a heart attack following publication of a clinical trial demonstrating its effectiveness, then mortality should fall for heart-attack patients treated at all hospitals. If innovation is low-tech, but knowledge barriers exist, then diffusion of the innovation will take some time. In this case, an outcome gap will develop between hospitals that "lead" and those that "trail" in the diffusion of new knowledge about readily available treatments. The leading hospitals are likely to be those that treat high volumes of patients, so that their medical staff is more knowledgeable and specialized. If innovation is high-tech, it will tend to be used first at the hospitals that tend to adopt high-tech innovations early. Such hospitals might include those with easier access to capital or with medical staffs more oriented toward high technology. Patients treated at slower-adopting hospitals will tend to be treated with the technology

later, if at all. Finally, teaching hospitals (which conduct the bulk of medical research) are likely to be the first to introduce technologies that are truly "new."

Differential adoption of each type of innovation will thus result in differences in changes in outcomes or expenditures for hospitals that differ in patient volume, proclivity for adopting high-tech treatments early, and teaching status. Over more than two periods, "catching up" can occur. For example, patients treated by small, late-adopting hospitals may gain access to low-tech innovations as knowledge diffuses, and to high-tech innovations as a result of transfers or readmissions to high-tech hospitals as the patients and the physicians learn about the previously adopted innovations. Unless innovation continues at the higher-volume, higher-tech hospitals, any differences from the previous period may be eliminated.

Table 1 reviews some results of our comparisons of trends in heart-attack outcomes across different types of hospitals between 1984 and 1994. The table divides technological change into two broad periods, 1984–1991 and 1992–1994, reflecting the more rapid aggregate expenditure growth that has occurred in recent years. For each outcome, two sets of estimates are reported: the average annual rate of change in outcomes and expenditures for the 1984–1991 period, and the *difference* in this trend for the subsequent 1992–1994 period. The first row of the table (panel A) reports the average annual rates of change for all hospitals and heart-attack patients. Like the preceding figures, it shows that one-year mortality declined substantially throughout the entire study period, by an average of one percentage point per year, while the rate of subsequent complications from heart disease remained approximately unchanged. Expenditures increased throughout the study period, but especially after 1991. We focus here on hospital expenditures; our studies of total expenditure growth show that hospital expenditures comprise the bulk of heart-attack expenditures and expenditure growth.

The lower rows of the table (panel B) show the difference in trends for each specialized hospital type (teaching, early adopter of high-tech procedures as indexed by catheterization adoption, and heart-attack

TABLE 1—CHANGES IN OUTCOMES AND EXPENDITURES FOR ELDERLY HEART-ATTACK PATIENTS

	One-year mortality		One-year hospital expenditure	
	Average annual trend, 1984–1991	Trend difference, 1992–1994	Average annual trend, 1984–1991	Trend difference, 1992–1994
A. All Hospitals:				
Overall adjusted trend	–1.03 (0.01)	0.03 (0.05)	548 (4)	594 (14)
B. Trends by Hospital Types:				
Base ^a	–0.75 (0.05)	–0.42 (0.17)	428 (14)	532 (51)
Differences from base:				
Major teaching hospital	0.03 (0.05)	–0.23 (0.17)	157 (16)	225 (51)
Adopted catheterization by 1984	–0.13 (0.04)	0.03 (0.14)	–214 (13)	139 (41)
Adopted catheterization 1985–1991	–0.17 (0.04)	0.29 (0.14)	–51 (13)	–76 (42)
Adopted catheterization 1992–1994	–0.001 (0.067)	–0.32 (0.21)	24 (20)	233 (64)
Intermediate heart-attack volume ^b	–0.16 (0.06)	–0.36 (0.19)	–151 (17)	84 (58)
High heart-attack volume ^c	–0.22 (0.06)	0.46 (0.21)	268 (18)	–113 (62)

Note: Numbers in parentheses are standard errors.

^a Nonteaching, low-tech, low-volume hospitals.

^b From 25 to 99 patients/year.

^c More than 100 patients/year.

volume) relative to a base trend for nonteaching, low-intensity, low-volume hospitals. Trends in outcomes and expenditures differed among the hospital types and also differed between the 1984–1991 and 1992–1994 periods. Between 1984 and 1991, though mortality improved to some extent at all hospital types, greater improvements occurred for the hospitals that tended to adopt high-tech procedures earlier. Hospitals that had adopted catheterization by 1984, or that adopted between 1985 and 1991, had average annual improvements in mortality that were 0.13–0.17 percentage points greater than the base trend. The highest-volume hospitals also had somewhat greater annual rates of improvement (additional decline of around 0.2 percentage points per year). In contrast, major teaching

hospitals had no substantial additional mortality improvements. These results are consistent with an additional effect of high-tech innovations, leading to greater mortality gains for the hospitals that tended to adopt high-tech procedures early, and an increasing “knowledge gap,” leading to greater mortality gains at the higher-volume hospitals.

In contrast, the trends for 1992–1994 suggest substantially different consequences of technological innovation. The base trend in mortality for patients first treated by the small, lower-tech hospitals was much more negative (an average annual decline of 1.17 percentage points, 0.42 percentage points more than for 1984–1991). Hospitals that tended to adopt high-tech innovations early also showed gains in mortality improvements relative to the 1984–1991 period, but trend differences were not as great as for the small hospitals. For example, at hospitals that adopted catheterization between 1985 and 1991, the average annual mortality change between 1992 and 1994 ($-0.75 - 0.17 - 0.42 + 0.29 = -1.05$) was no longer significantly more negative than the base trend. For high-volume hospitals, the trend difference was even more dramatic (e.g., a differential of +0.46 for the highest-volume hospitals). Thus, mortality at the smallest hospitals was catching up with the gains at the hospitals with the greatest experience in heart-attack treatment. These results suggest that innovations such as drug treatments which could be used on patients at all hospitals were relatively important in explaining mortality improvements, and that the “knowledge gap” in applying treatments between more- and less-experienced hospitals was diminishing.

Table 1 also shows the differential trends in one-year hospital expenditures. Major teaching hospitals had relatively rapid expenditure growth, especially after 1991. Higher-volume hospitals also had more rapid expenditure growth throughout the study period, though the differentials did not become significantly larger after 1991. The more rapid growth for higher-volume hospitals was primarily the result of increased complications in heart-attack survivors, leading to more expenditures for later hospital readmissions. Perhaps surprisingly, hospitals that were more likely to adopt

high-tech procedures did not consistently have greater expenditure growth. Though use of more high-tech procedures led to greater expenditure growth during the initial hospital admission, patients treated by these hospitals had relative declines in later hospital admissions, offsetting the more rapid growth in initial treatment expenditures.

We have estimated similar models of longer-term mortality and expenditure outcomes, as well as complications related to quality of life. With reasonable valuations for the improvements in survival and changes in expected quality of life, the mortality improvements from 1984 to 1991 appear to be clearly worthwhile on average, with notable welfare gains associated with high-tech innovation. From 1992 to 1994, the more rapid growth in expenditures resulted in more modest aggregate welfare gains. Moreover, the bulk of these gains occurred at small hospitals and hospitals that developed high-tech capabilities later; high-volume hospitals and hospitals adopting high-tech procedures early generally became less productive. Collectively, the results suggest that high-tech innovations and innovations requiring specialized knowledge and experience were important components of the improvement in heart-attack mortality between 1984 and 1991. But these types of technological change were relatively less valuable, and potentially excessive, in more recent years.

III. Conclusion

Understanding the welfare consequences of technological change is a critical issue for designing optimal health-care policies. Even if the expenditure growth resulting from technological change has been worthwhile on average, many medical technologies may be used excessively, inadequately, or inappropriately. Understanding which types of medical technologies diffuse efficiently, which do not, and what policies influence diffusion is crucial for evaluating productivity changes in this enormous sector of the economy. Expenditure growth and outcome improvements have differed substantially across different types of hospitals over time. Exploring these differences is likely to provide another source of insights into the nature, determinants, and consequences of

technological change in health care. Future research on medical productivity is likely to continue to integrate detailed analysis of medical treatment decisions with the analysis of aggregate outcome and expenditure effects.

REFERENCES

- Braunwald, Eugene. "Shattuck Lecture—Cardiovascular Medicine at the Turn of the Millenium: Triumphs, Concerns, and Opportunities." *New England Journal of Medicine*, 6 November 1997, 337(19), pp. 1360–69.
- Cutler, David and McClellan, Mark. "Technological Change in Medicare," in David Wise, ed., *Topics in the economics of aging*. Chicago: University of Chicago Press, 1998 (forthcoming).
- Cutler, David; McClellan, Mark; Newhouse, Joseph and Remler, Dahlia. "Are Medical Prices Declining?" National Bureau of Economic Research (Cambridge, MA) Working Paper No. 5750, 1996.
- Gelijns, Annetine and Rosenberg, Nathan. "The Dynamics of Technological Change in Medicine." *Health Affairs*, Summer 1994, 13(3), pp. 28–46.
- Heidenreich, Paul and McClellan, Mark. "The Consequences of Technological Change in Heart Attack Care, 1975–1995: A Literature Review and Synthesis." Mimeo, Stanford University, 1998.
- Hunink, Maria; Goldman, Lee; Tosteson, Anna; Mittleman, Murray; Goldman, Paula; Williams, Lawrence; Tsevat, Joel and Weinstein, Milton. "The Recent Decline in Mortality from Coronary Heart Disease, 1980–1990." *Journal of the American Medical Association*, 19 February 1997, 277(7), pp. 535–42.
- ISIS-2 (Second International Study of Infarct Survival) Collaborative Group. "Randomized Trial of Intravenous Streptokinase, Oral Aspirin, Both, or Neither Among 17,187 Cases of Suspected Acute Myocardial Infarction." *Lancet*, 1988, 2(8607), pp. 349–60.
- Krumholz, Harlan; Murillo, Jaime; Chen, Jersey; Vaccarino, Viola; Radford, Martha; Ellerbeck, Edward and Wang, Yun. "Thrombolytic Therapy for Eligible Elderly Patients with

- Acute Myocardial Infarction." *Journal of the American Medical Association*, 4 June 1997, 277(21), pp. 1683-88.
- Lewis, H. D.; Davis, J. W. and Archibald, D. G. "Protective Effects of Aspirin Against Acute Myocardial Infarction and Death in Men with Unstable Angina: Results of a Veterans Administration-Cooperative Study." *New England Journal of Medicine*, 18 August 1983, 309(7), pp. 396-403.
- McClellan, Mark. "Hospital Reimbursement Incentives: An Empirical Approach." *Journal of Economics and Management Strategy*, Spring 1997, 6(1), pp. 131-65.
- McClellan, Mark; McNeil, Barbara and Newhouse, Joseph. "Does More Intensive Treatment of Acute Myocardial Infarction Reduce Mortality?" *Journal of the American Medical Association*, 21 September 1994, 272(11), pp. 1470-77.

The Value of Health: 1970–1990

By DAVID M. CUTLER AND ELIZABETH RICHARDSON*

Given a choice between spending more money on medical care or on other consumption goods, which should society choose? Should the National Institutes of Health devote a larger part of its research budget to AIDS or to cancer? Has the increased inequality of income in the United States led to worse health among the poor? Answering these questions is fundamental to understanding the medical sector and to forming sound public policy. But knowledge about the value of health is limited. At both a conceptual and empirical level, there are few integrated treatments of population health. In this paper and related work (Cutler and Richardson, 1997), we estimate the health of the U.S. population and examine how it has changed over the past several decades.

I. Methodology

“Health” is a multi-attribute concept, encompassing both physical and mental components. The first requirement in measuring health is to find a way to combine these different dimensions. We assume that a person’s quality of life in any year (denoted H_t) can be scaled on a 0 to 1 basis, where 0 is death and 1 is perfect health. Living with a given disease falls between 0 and 1. We can then add up the expected number of quality-adjusted life years a person has remaining, to form a measure of expected quality-adjusted life years, or QALY’s. A QALY has a unit of years in perfect health; if we multiply the number of QALY’s by the value of a year in perfect health (denoted V), we have a measure of the value of health. Following Michael Grossman’s (1972) pioneering work, we term

this measure “health capital.” Algebraically, health capital is defined as

$$(1) \quad (\text{Health Capital})_t = V \sum_{k=0}^{\infty} \frac{E_t[H_{t+k}]}{(1+r)^k}$$

where r is the real discount rate. Health capital is analogous to the more common measure of human capital in the economics literature.

Measuring health capital empirically requires assuming values for V and r . Both of these are longstanding issues in economics. A consensus estimate from the literature (W. Kip Viscusi, 1993; George Tolley et al., 1994) is that a life is worth about \$3 million–\$7 million, or that a life year is worth about \$75,000–\$150,000. We use an intermediate value of \$100,000 for a year of life. The economics literature (Edward M. Gramlich, 1990) also suggests a real discount rate of about 3 percent, which we employ.

Measuring quality of life is more difficult. We start with the probability that the person is alive or dead in each year in the future. Estimates of survival probabilities conditional on reaching any age are published in standard life tables.¹ Life expectancy has increased over time, from 47 years at birth in 1900, to 71 years in 1970, to 75 years in 1990. Increased survival implies increasing health over time.

We then adjust these survival rates by the prevalence of disease at every age. We measure disease prevalence using the annual National Health Interview Surveys (NHIS). The NHIS asks about a variety of acute and chronic conditions. After reviewing the NHIS documentation and other sources of data, we identified 10 conditions that we believe to be

* Cutler: Department of Economics, Harvard University, Cambridge, MA 02138, and National Bureau of Economic Research; Richardson: Department of Health Care Policy, Harvard University, 180 Longwood Avenue, Boston, MA 02115. We are grateful to Victor Fuchs, Theodore Keeler, and Doug Staiger for helpful comments.

¹ Note that this measure is not perfect. Period life tables assume that people currently alive will have age-specific mortality in the future equal to current mortality at those ages. Any improvements in mortality that can be forecast are thus omitted from the life table. There are no better alternatives to the period life table, however, so we use these data.

TABLE 1—DISEASE INCIDENCE AND QUALITY OF LIFE

Condition	Prevalence		QALY weight	
	1970	1990	1970	1990
Amputee	6.1	6.0	0.87	0.89
Arthritis ^a	111.8	127.8	0.69	0.79
Blindness	8.6	2.0	0.73	0.87
Other vision	48.0	30.2	0.84	0.93
Cancer ^b	11.1	18.7	0.70	0.70
Cardiovascular disease ^a	64.7	99.3	0.57	0.71
Diabetes ^a	45.9	54.3	0.65	0.66
Hearing	80.9	91.2	0.91	0.93
Orthopedic ^a	102.1	135.0	0.70	0.88
Paralysis ^a	7.4	7.1	0.62	0.68

Note: Prevalence is adjusted for the change in the age- and sex-mix of the population.

^a There are also interactions for these QALY estimates which are not reported.

^b QALY estimate is based on review of literature rather than model estimate.

consistently reported between 1970 and 1990. The conditions and their prevalence (adjusted for the changing age and sex mix of the population) in 1970 and 1990 are shown in the first columns of Table 1. The prevalence of most of these conditions is increasing over time. Between 1970 and 1990, for example, the four most common conditions (orthopedic problems, arthritis, cardiovascular disease, and hearing problems) all increased in prevalence, by up to 50 percent. This increase in disease prevalence is in opposite direction to the longevity improvement.

Finally, we need to attach quality-of-life weights to each condition. Quality of life differs across conditions and over time; the fact that more buildings have ramps and elevators, for example, raises quality of life for those with mobility problems. The most common method for estimating quality weights is through surveys (George W. Torrance, 1986). For example, people might be asked how many years of perfect health they would trade off for a given number of years with a particular condition (termed the "time trade-off" method). The answer to this question gives an implicit quality adjustment for the disease. In practice, however, there is no consensus in the literature about the disutility associated with various conditions or the change in these disutilities over time.

We therefore follow an alternative approach to quality measurement. The Health Interview Survey asks people to rate their health as either excellent, very good, good, fair, or poor.² We assume that a person's underlying health, h_i^* , is related to demographics and health conditions (X_i) as

$$(2) \quad h_i^* = X_i\beta + \varepsilon_i.$$

If we assume that peoples' self-reported health reflects their underlying health state, we can estimate the β coefficients using the self-report data. In particular, if ε is normally distributed, equation (2) can be estimated as an ordered probit model for self-reported health. The β 's then give the reduction in quality of life associated with each condition.³

The last columns of Table 1 show the quality-of-life weight for each condition in 1970 and 1990. The estimates generally accord with intuition: quality of life is lowest for cancer, cardiovascular disease, and diabetes, and highest for minor vision and hearing problems. Importantly, quality of life for each condition is improving over time. Whether because of medical care or other factors, people consistently report themselves in less worse health than they did in the past. The reduced severity of disease over time implies improved health.

Combining our estimates of the share of people who are alive, the prevalence of people with particular conditions, and the quality of life for people with those conditions, we can estimate quality of life as

$$(3) \quad H_{t+k} = \Pr[\text{Alive at } t+k] \\ \times \left(\sum_d \Pr[\text{Condition } d \text{ at } t+k] \right) \\ \times [\text{QALY for } d \text{ at } t+k]$$

² Prior to 1982, very good was omitted.

³ For a more detailed discussion of this issue, see Cutler and Richardson (1997).

TABLE 2—HEALTH CAPITAL

Measure	Health capital (thousands of dollars)			Change, 1970–1990
	1970	1980	1990	
By age:				
0	2,350	2,395	2,444	95
65	590	700	759	169
By race and sex, age 0:				
Men				
White	2,364	2,404	2,471	108
Black	2,255	2,348	2,377	132
Women				
White	2,348	2,387	2,437	89
Black	2,226	2,305	2,361	135
By income:				
Age 0				
Below poverty level	2,240	2,285	2,296	57
Above poverty level	2,361	2,365	2,420	54
Age 65				
Below poverty level	567	706	735	168
Above poverty level	611	707	756	145

Notes: The value of a year in perfect health is assumed to be \$100,000, and the real discount rate is assumed to be 3 percent.

where d is the range of conditions a person may have.

II. Trends in Health

Table 2 presents estimates of health capital in 1970, 1980, and 1990. Health capital is large. For a newborn in 1990, health capital over the lifetime is about \$2.5 million; for a person aged 65, remaining health capital is about \$750,000. By comparison, if a person earns \$30,000 per year from ages 20 through 65, the present discounted value of lifetime earnings (as of age 0) is less than \$450,000. Of course, there is no requirement that people be able to afford the value they place on their health.⁴

Perhaps more importantly, health has improved over time. For newborns, health capital

increased by \$95,000 between 1970 and 1990, while health capital for the elderly increased by \$169,000. The greater increase in health capital for the elderly than for the young is a result of differential changes in mortality by age. The lion's share of mortality reduction between 1970 and 1990 was a result of fewer deaths from cardiovascular disease. Since cardiovascular disease is more prominent late in life than early in life, the present value of these gains is greater for the elderly than for the young.

Are these changes in health large or small? One way to gauge them is to compare the increase in health capital with the increase in medical spending over this time period. While not all of the increase in health capital results from improved medical care, if we find that the increase in health capital is smaller than the increase in medical spending, that would be a strong indication that medical spending over this time period was not, on net, worth it. Using cross-section data on medical spending in 1970 and 1987, we estimate that expected medical costs increased by \$19,000 for infants (in 1987 dollars), and by \$34,000 for people aged 65. The increase in health capital is greater than the increase in medical spending; thus, the return to medical care could be very high.

The next rows of Table 2 show health capital by race and sex. We show health capital just for newborns. Changes in health capital by race mirror changes in labor-market earnings by race. In the 1970's, black men's health improved relative to white men; the same is true of earnings. But this situation reversed in the 1980's. In that decade, health capital for black males rose by much less than for white males. This is also true of earnings.

Changes in health capital by gender are less reflective of changes in labor-market returns by gender. Health capital for men and women is about the same at birth and is much greater for women at older ages, where earnings are much lower for women than for men. Further, health capital for the two groups changed about the same over this period. Earnings for women, in contrast, increased much more rapidly than did earnings for men in the 1980's. Understanding why changes in health capital by gender look so different from changes in

⁴ This is particularly true since estimates of the value of life, or of a life year, generally consider the value of small changes in mortality, so that wealth constraints are minimized.

economic returns by gender is an important topic of future research.

Measuring changes in health capital by income or education is more difficult than measuring changes by race and by gender, in large part because mortality is rarely classified by economic status. The lower panel of Table 2 shows estimates of health capital by income, measured by poverty status. Health capital is higher for people with family income above poverty level. A newborn in a family with income above poverty level has about \$124,000 more health capital than one in a family with income below poverty level; while a 65-year-old above poverty level has about \$21,000 more health capital than a 65-year-old with lower income.

The increase in health capital from 1970 to 1990 was about the same for both income groups. However, the gap in health for persons over 65 narrowed over this period, with elderly individuals below the poverty level realizing larger gains in health capital than elderly individuals above the poverty level.

III. Conclusions

The U.S. population is healthier than it used to be. We estimate that health improved by \$100,000–\$200,000 per person between 1970 and 1990. This increase in health is greater for the elderly than for the young. It is also greater than the increase in medical spending over this time period, although medical spending certainly did not cause all of the improved health. Health changes for blacks mirror labor-market changes for that group, but we do not find improvements for women relative to men. The

gap in health inequality by income has declined for people over 65.

Perhaps most important, our measure of health capital will allow us to examine the value of the medical-care system more systematically. We know a great deal about the resources we put into the medical system—people, dollars, technology, and the like—but very little about what we get out. Our methodology provides a mechanism for estimating what return we get for our medical-care dollars.

REFERENCES

- Cutler, David M. and Richardson, Elizabeth. "Measuring the Health of the United States Population." *Brookings Papers on Economic Activity, Microeconomics* 1997, pp. 217–71.
- Gramlich, Edward M. *A guide to benefit–cost analysis*, 2nd Ed. Englewood Cliffs, NJ: Prentice Hall, 1990.
- Grossman, Michael. *The demand for health: A theoretical and empirical investigation*. New York: Columbia University Press, 1972.
- Tolley, George; Kenkel, Donald and Fabian, Robert, eds. *Valuing health for policy: An economic approach*. Chicago: University of Chicago Press, 1994.
- Torrance, George W. "Measurement of Health State Utilities for Economic Appraisal." *Journal of Health Economics*, March 1986, 5(1), pp. 1–30.
- Viscusi, W. Kip. "The Value of Risks to Life and Health." *Journal of Economic Literature*, October 1993, 31(4), pp. 1912–46.

Economic Effects of Reducing Disability

By KENNETH G. MANTON, ERIC STALLARD, AND LARRY CORDER*

Assessing the impact of chronic disability on the U.S. economy in general, and health expenditures in particular, involves a number of complex data and analytic problems. One involves measurement. Disability varies over individuals, both in content (i.e., there are multiple dimensions of physical and cognitive functioning) and intensity. Moreover, in contrast to most chronic diseases, there is no well established metric or "staging" of chronic disability. One solution to the measurement problem has been to discount life expectancy subjectively by the perceived degree of functional loss manifest at specific ages. However, such subjective evaluations of quality of life have the problem that they vary tremendously depending upon the values and perspectives of the evaluator (e.g., Medicare, health-care providers, the patient) or, even more difficult, of the same evaluator in performing different social and economic roles (e.g., a health-system administrator who evaluates benefits in one way for a plan he manages and in a different way when, as a consumer, he is evaluating the benefits of an expensive treatment for a member of his family).

A potentially more fruitful, and objective, approach uses psychometric methods where latent traits, like disability, are assumed to be inferable from a wide battery of measures where the covariation/codependency of those measures is used to identify specific dimensions of disability (Manton et al., 1991). Effectively, it is assumed that the state variables underlying measurements can be identified from the measures by determining the probabilistic properties of the state variables necessary to make the observed measures independent, conditional upon the analytically inferred state variable distribution (e.g., P. Suppes and M. Zanotti, 1981). However, the problem of establishing an appropriate dis-

ability metric remains in standard multivariate models based on the Gaussian distribution, where one is limited to the information contained in the distribution's first two sets of statistical moments. One can do more with multivariate non-Gaussian distributions, where there are explicit functional relations between location and scale parameters containing additional information.

Thus, a possible solution is to use specialized multivariate procedures which preserve the constraints imposed by the metric of the original discrete measurements. This conservation is difficult for continuous variables assessed on different metrics but can be done using the Grade of Membership model (Manton et al., 1994) applied to multidimensional categorical data where convexity constraints on the model parameter space restrict their maximum-likelihood estimates to fall within the range spanned by the original measures (H. Weyl, 1949). The use of convexly constrained disability parameters also conserves the duration metric of life expectancy so that it can be additively decomposed into the time expected to be spent in each analytically determined disability state or partition (Max A. Woodbury et al., 1997).

Using duration-weighted measures of functional status more directly reflects the amount of human capital preserved (or generated) in a population by a health intervention. If K is the number of basic functional states, M the number of disability measures, T time, a age, and i indexes individuals ($i = 1, 2, \dots, I$), then the disability parameters involve numerically evaluating a multidimensional stochastic integral over a high-dimensional $(M \times K) \times (T \times a) \times I$ space (H. Dennis Tolley and Manton, 1992). The integral can be disaggregated as is needed, for example, to identify changes in human capital for specific birth cohorts or time periods. In this framework, diseases represent markers of specific pathological changes that alter a duration-weighted human-capital index. The index combines the

* Center for Demographic Studies, Duke University, 2117 Campus Drive, Durham, NC 27708-0408.

effects of chronic functional and mortality (life-length) changes so that both the quality and quantity of life are numerically expressed. An advantage of using a common disability metric is that intervention effects can be compared across diseases, interventions, and time. Time is of particular concern analytically in that recent, let alone future, health and biomedical changes have occurred too fast for us to have the fully mature experience of the interaction of each person's health history with recent health-care innovations (e.g., a person first eligible for Medicare in 1982 cannot be more than 80 years of age in 1997).

Though this index reflects both physical and cognitive impairments, and their evolution through time, an issue still to be resolved is how to value objectively each unit of human capital (i.e., a multiplier of the human capital unit has to be estimated for each person to determine economic value). This multiplier is affected by a number of factors. It is likely to change as our technical ability to reduce chronic disability increases (i.e., the emergence of technologies to avoid disability are also likely to produce technologies to physically mitigate the effects of disability on an individual) (Manton et al., 1993a; Manton and Stallard, 1995). Thus, technological advances are likely to interact to both produce overall increases in human capital and increases in the economic value of human capital among still disabled persons.

To translate this into empirical terms we analyzed 27 measures of activities of daily living (ADL), instrumental activities of daily living (IADL), and physical impairments made in the 1982, 1984, 1989, and 1994 National Long Term Care Survey (NLTCs) (Manton et al., 1997b). Analyses were done with cross-temporal constraints imposed on the coefficients defining the K state variables from the J measures so that the individual scores on the state variables were defined comparably at each time point. Thus, the $g_{ik}(t)$ could be used to define partial trajectories of disability change for individuals. Since the data cover ages from 65 to roughly 115 years for the period 1982–1996 (1995 and 1996 mortality data are available from Medicare), there is partial information on cohort effects as well as health changes over age and period (Manton et al., 1997b). Estimates of disability scores

TABLE 1—LIFE EXPECTANCIES (e) AND MEAN GRADE OF MEMBERSHIP (GoM) SCORES (g) BY PURE TYPE (k) FOR AGES 65 AND 72

Pure type (k)	Age 65		Age 72	
	$e_k(x)$	$\bar{g}_k(x)$	$e_k(x)$	$\bar{g}_k(x)$
A. Men	15.74		10.98	
1	13.64	94.49	9.02	90.40
2	0.36	0.89	0.35	1.44
3	0.20	1.21	0.15	1.01
4	0.29	1.04	0.27	1.15
5	0.44	1.05	0.38	3.39
6	0.37	1.10	0.35	1.57
Institutional	0.42	0.22	0.47	1.04
B. Women	22.24		15.85	
1	15.71	94.87	10.31	87.62
2	1.54	0.99	1.56	2.31
3	0.60	1.01	0.56	1.58
4	0.75	0.40	0.77	1.14
5	1.26	1.10	1.19	4.69
6	0.81	0.76	0.82	1.38
Institutional	1.55	0.86	1.64	1.27

Notes: Life expectancies are reported in years; GoM scores are percentages.

and life expectancy are calculated on a cohort specific basis in the model. Thus, differential equations describing human capital processes reflect cohort improvements and produce higher life expectancies than those calculated from cross sectional data (Manton and Stallard, 1996; Manton et al., 1997b).

In Table 1 we present, for ages 65 and 72, the distribution of the time (in years) expected to be lived in each of seven functional states. Two different vectors are given for each age for males and females. The $\bar{g}_k(x)$ are the age-specific average scores for each of the seven functional states. The weighted prevalence (where g_{ik} 's are multiplied by sample weights w_i before averaging) of the most active type remains high for both males (90.4) and females (87.6), even to age 72. The second vector indicates the number of years expected to be lived in a health state after that age. From ages 65 to 72 much of the life span in active states is preserved. For males the time lived in the three most active states only declines from 88.0 percent to 86.6 percent. Female declines are larger because of the larger amount of time

females expect to live in institutions (e.g., 7.0 percent at age 65, 9.7 percent at age 72).

To translate the quantities in Table 1 into economic terms, several steps are required. First, the savings attributed to moving from a more disabled state to a less disabled state need to be assessed. This is difficult because of several factors. There is nothing in Medicare that ensures that funds are spent in an "optimal" way to maximize health improvements. Indeed, the medical literature suggests that, for leading clinicians, there is a 5–10-year lag between demonstrating the efficacy of a new treatment in clinical trials and its acceptance as a standard treatment in medical texts (E. M. Antman et al., 1992). Thus, using average Medicare or other medical expenditures for each type does not tell us what could, or should, be saved (or spent) given the best extant medical technology. It gives the national average of expenditures for the prevailing U.S. medical and health-care reimbursement practices, which will generally be suboptimal to the best demonstrated in recent trials. This also does not determine what could be saved if cost reductions were a factor assessed in a formal utility-optimization problem.

Confounding efforts to estimate the true costs and savings due to changes in disability is the problem of ascertaining the contribution to net economic activity of an additional year of life above age 65. Obviously, an evaluation is easy for a person who is either fully employed or who is generating a majority of maintenance income from assets and investments. These may partly offset Social Security expenditures. The offset will change as new elderly cohorts born in 1930 or later (and who have worked wholly in a post-Depression era) pass age 65 because of higher average levels of savings and assets.

To illustrate some of the costs and savings associated with changes in disability prevalence, consider the implications of the differences between males and females in the time expected to be lived in institutions. Assuming that institutionalization in 1996 costs \$51,356 per year, the difference in undiscounted lifetime costs above age 65 between males and females is \$59,612. To assess direct medical expenses is more difficult because Medicare benefits and Medicare program structure are

changing. Estimates from the 1987 National Medical Expenditure Survey (NMES) suggest that, for persons aged 65 and above, the ratio of expenditures for disabled persons to that for nondisabled persons is 2.75 for all but institutional care. This is using a very broad, inclusive definition of disability that may understate the ratio (W. Max et al., 1996). The ratio is much larger if one includes nursing-home costs. Their inclusion in 1987 implies an expenditure ratio of 6.3. These ratios would also likely be higher if more precise definitions of disability were used. For example, in Manton et al. (1993c), ratios of 5.5–7.0 were obtained for a multivariate index of disability. That index preserved measures of the population burden of disability because of convexity constraints on parameters. The issue of how costs and savings change with the distribution of a given population's burden of disability is an important analytical problem (e.g., Tolley and Manton, 1998).

Since these ratios incorporate all real dollar expenditures on health they must also include expenditures on the medical therapies that reduced disability for individuals. The year 1987 is the midpoint of a period (1984–1989) over which significant disability declines were observed in the United States (Manton et al., 1993b, 1997a). Cost estimates may involve medical expenditures calculated either using actual expenditures or using ideal expenditures based on an expert judgment of what the ideal treatment, and its cost, should be. In some cases the ideal treatment is a much less expensive treatment. This is the case for gastric ulcer, where antibiotic therapy may cure many cases of gastric ulcer by eradicating *Helicobacter pylori* from the gastric mucosa.

In Table 2, we consider the effects of disability on the population whose average age at retirement is 65 (or 72). The economic unit is the value (\$1) of an annuity where expenditures are discounted over time at 4 percent per year. For example, for men, increasing retirement age to 72 would save \$5.58 in annuity costs. The costs of raising the retirement age are the costs of supporting additional persons who remain in a chronic disability state. Thus, for men, the net gain at age 72 is $\$5.58 - 0.22 = \5.36 . If one were being paid \$10,000 per year, this would imply a current value of

TABLE 2—PRESENT VALUE (AT AGE 65) OF AN ANNUITY
PAYING \$1 ANNUALLY UNTIL RETIREMENT
IF RETIREMENT WERE DELAYED AFTER AGE 65

	Men		Women	
	Age 65.5	Age 72.0	Age 65.5	Age 72.0
Annuity savings by extending working cap (\$) ^a	0.4933	5.5752	0.4945	5.8291
Costs of disabled (\$)	0.0136	0.2201	0.0212	0.3302
Ratio	97.2	96.1	95.7	94.3
Disability groups: ^b				
3) IADL impairments and physical limitations	0.0030	0.0357	0.0046	0.0653
4) IADL impairments, some early cognitive impairments	0.0025	0.0392	0.0034	0.0510
5) Physical impairments and mobility limitations	0.0044	0.0675	0.0059	0.0917
6) Frail	0.0037	0.0515	0.0035	0.0563
7) Institutional care	0.0000	0.0262	0.0038	0.0659

^a Annuity pays while individual is in the state.

^b Disability states 1 (no ADL or IADL disability) and 2 (no disability or physical limitations) are not included. The disability groups characterize subsets of the population by their dominant disability traits from a vector of 27 traits.

\$53,600 to be saved by the Social Security Administration for a person who continued to work from age 65 to 72. For a cohort of a million persons, this would be \$53.6 billion in current dollars.

REFERENCES

- Antman, E. M.; Lau, J.; Kupelnick, B.; Mosteller, F. and Chalmers, T. C. "A Comparison of Results of Meta-analyses of Randomized Control Trials and Recommendations of Clinical Experts: Treatments for Myocardial Infarction." *Journal of the American Medical Association*, July 1992, 268(2), pp. 240-48.
- Manton, Kenneth G.; Corder, Larry S. and Stallard Eric. "Changes in the Use of Personal Assistance and Special Equipment 1982 to 1989: Results from the 1982 and 1989 NLTCs." *Gerontologist*, 1993a, 33(2), pp. 168-76.
- _____. "Estimates of Change in Chronic Disability and Institutional Incidence and Prevalence Rates in the U.S. Elderly Population from the 1982, 1984, and 1989 National Long Term Care Survey." *Journal of Gerontology Social Sciences*, 1993b, 47(4), pp. S153-66.
- _____. "Chronic Disability Trends in the U.S. Elderly Populations 1982 to 1994." *Proceedings of the National Academy of Sciences*, March 1997a, 94, pp. 2593-98.
- Manton, K. G.; Singer, B. H. and Suzman, R. M., eds. *Forecasting the health of elderly populations*. New York: Springer-Verlag, 1993c.
- Manton, Kenneth G. and Stallard, Eric. "Change in Health, Mortality, and Disability and the Impact on Long Term Care," in M. E. Cowart and J. Quadagno, eds., *Long term care: Conference proceedings*. Tallahassee, FL: Pepper Foundation, 1995.
- _____. "Longevity in the United States: Age and Sex Specific Evidence on Life Span Limits From Mortality Patterns: 1960-1990." *Journal of Gerontology: Biological Sciences*, 1996, 51(5), pp. B362-75.
- Manton, Kenneth G.; Stallard, Eric and Corder, Larry S. "Changes in the Age Dependence of Mortality and Disability: Cohort and Other Determinants." *Demography*, February 1997b, 34(1), pp. 135-57.
- Manton, Kenneth G.; Woodbury, Max A. and Stallard, Eric. "Statistical and Measurement Issues in Assessing the Welfare Status of Aged Individuals and Populations." *Journal of Econometrics*, October-November 1991, 50(1-2), pp. 151-81.
- Manton, Kenneth G.; Woodbury, Max A. and Tolley, H. Dennis. *Statistical applications using fuzzy sets*. New York: Wiley, 1994.
- Max, W.; Rice, D. P. and Trupin, L. "Medical Expenditures for People with Disabilities." *Disability Statistics Abstract*, December 1996, 12(12), pp. 1-4.
- Suppes, P. and Zanotti, M. "When Are Probabilistic Explanations Possible?" *Syntheses*, January 1981, 48, pp. 191-99.
- Tolley, H. Dennis and Manton, Kenneth G. "Large Sample Properties of Estimates of Discrete Grade of Membership Model." *Annals of Statistical Mathematics*, 1992, 44(1), pp. 85-95.
- _____. "Mortality Models with Health State Variables." Unpublished manuscript, Center for Demographic Studies, Duke University, 1998.

Weyl, H. "The Elementary Theory of Convex Polyhedra," in H. Kuhn and A. Tucker, eds., *Contributions to the theory of games*. Princeton, NJ: Princeton University Press, 1949, pp. 3-18.

Woodbury, Max A.; Manton, Kenneth G. and Tolley, H. Dennis. "Convex Models of High Dimensional Discrete Data." *Annals of Statistical Mathematics*, 1997, 49(2), pp. 1-23.

Measuring Prices and Quantities of Treatment for Depression

By RICHARD G. FRANK, SUSAN H. BUSCH, AND ERNST R. BERNDT*

The assessment of productivity in the health sector relies on applying price indexes to expenditure data (Jack Triplett, 1998). Changes in "output" are calculated indirectly by deflating expenditure increases by a price index. Choosing an appropriate price index will therefore be central to how changes in expenditures are interpreted for policy purposes. In this paper we propose an approach to constructing a price index for the treatment of an important chronic illness, major depression. Major depression affects about 10 percent of the U.S. population in a 12-month period and accounts for over \$20 billion in medical expenditures each year. Moreover, spending on depression and other mental illnesses, during the late 1980's and early 1990's, has been pointed to as a factor driving overall medical spending upward. Sorting out whether spending increases are due to price increases, expanded use of marginally effective services, or expanding productivity is critical for making policy judgments in the health sector. Our approach to constructing a price index builds on previous research by attempting to define quantities that better approximate what individuals seek from spending on medical care, effective treatment for an episode of illness (Zvi Griliches, 1962; Anne Scitovsky, 1967; CPI Commission, 1996). However, because we cannot directly observe effectiveness of treatment from data based on

insurance claims and spending, we use indirect methods to identify effective bundles of treatment for major depression.

We make use of data on transaction prices and the actual out-of-pocket payments made by consumers to estimate the supply and demand prices of bundles of effective treatment for depression. The data are derived from the experiences of a large insured population for the years 1991–1995. During this period, important new treatments for depression were diffusing through the health sector, accompanied by dramatic changes in how care is rationed in the form of managed care. The data we use reflect these trends and are incorporated into the price-index calculations. Finally, we compare four formulations of demand-price and supply-price indexes that make differing assumptions *ex ante* about substitution among bundles of treatment.

I. Definition and Measurement of the Price for the Treatment of Depression

Depression is commonly characterized by melancholy, diminished interest and pleasure in all or most activities, and feelings of worthlessness. The clinical definition of major depression provides a very specific set of clinical criteria that must be met in order for a patient's condition to be considered as an episode of major depression. Specifically, the *Diagnostic and Statistical Manual of the American Psychiatric Association*, 4th Edition (DSM-IV), defines major depression as:

The presence of one of the first two symptoms, as well as at least five of nine total symptoms. The symptoms must be present most of the day almost every day, for at least two weeks. The symptoms include:

- 1) depressed mood most of the day nearly every day;
- 2) markedly diminished interest or pleasure in almost all activities most of the day;

* Frank and Busch: Harvard Medical School, Department of Health Care Policy, 180 Longwood Avenue, Boston, MA 02115; Berndt: Massachusetts Institute of Technology, Alfred P. Sloan School of Management, 50 Memorial Drive, E52-452, Cambridge, MA 02142. We gratefully acknowledge research support from Eli Lilly and Company, the U.S. Bureau of Economic Analysis, National Science Foundation Grant SBR-9511550, and National Institute of Mental Health Grant MH43703. We are also grateful to Douglas Cocks, Thomas Croghan, David Cutler, Dennis Fixler, Mark McClellan, Will Manning, Thomas McGuire, Joseph Newhouse, Charles Phelps, and Jack Triplett for helpful comments on an earlier draft. Elizabeth Notman provided key programming support for this project.

- 3) significant weight loss/gain;
- 4) insomnia/hypersomnia;
- 5) psychomotor agitation/retardation;
- 6) feelings of worthlessness (guilt);
- 7) fatigue;
- 8) impaired concentration (indecisiveness); and
- 9) recurrent thoughts of death or suicide

(American Psychiatric Association, 1994 p. 161).

It has been estimated that in the early 1990's 10.3 percent of the U.S. population met the criteria for major depression at sometime during a 12-month period (Ronald Kessler et al., 1994). The vast majority of individuals who experience an episode of major depression will return to their original level of functioning. However, between 20 percent and 35 percent experience persistent symptoms; these cases are commonly referred to as chronic depression. Furthermore, approximately 50 percent of all people having depressive episodes are expected to have a recurrence (American Psychiatric Association, 1993). Once an individual has a second episode, recurrence is 70 percent likely.

A. Alternative Treatments

In this research we focus on treatment for the acute phase of care. Research on the continuation phase of treatment is less developed, and definitive protocols have not been as widely adopted in many clinical settings. Treatments for major depression have advanced rapidly during the past 20 years. In the area of psychotherapy, various new techniques have expanded treatment options beyond psychodynamic or psychoanalytic approaches. Interpersonal therapy (IPT), behavior therapy (BT), family therapy, and cognitive behavior therapy (CBT) are all relatively new.

Extraordinary advances have been achieved in the area of antidepressant medication. Antidepressant medication has three general classes. These are (i) cyclic antidepressants, which include the widely used tricyclic antidepressants (TCA's) and a number of lesser-known drugs such as trazodone; (ii) selective serotonin reuptake inhibitors (SSRI's), which include brand-name drugs such as Prozac, Zo-

loft, Paxil, and Luvox; and (iii) monoamine oxidase (MAO) inhibitors, which, due to side effects and dangerous interactions, are generally used only for cases that are resistant to other forms of treatment. The newer SSRI's offer some distinct advantages over older TCA's. SSRI's are associated with lower risk of overdose, and fewer and lower levels of a number of side effects. Key side effects associated with TCA's include drowsiness, dry mouth, impaired ability to concentrate, seizures, and weight gain. SSRI's have been associated with side effects related to sexual dysfunction and anxiety. The advantages of SSRI's come at a significantly higher pecuniary cost than most TCA's. Psychotherapeutic interventions have been frequently combined with antidepressant medication as a strategy for treating major depression.

A typical population of patients consists of patients with varying levels of severity, and for given levels of severity, alternative treatments are provided. We develop a set of treatment "bundles" that group therapies in what we term therapeutically similar groups for treatment of a specific form of major depression. Our goal here is to identify treatment bundles that result in similar expected mental health outcomes. The implicit assumption we adopt is that obtaining similar outcomes from alternative treatments begins to approximate similar utility levels. We divide levels of depression into two classes: severe and less severe (hereafter, mild). In order to classify therapies into therapeutically similar treatment bundles, we have reviewed approximately 30 major clinical trials and meta-analyses from the clinical literature dealing with acute-phase treatment (Busch et al., 1996). This constitutes an indirect approach to incorporating outcome measures into the definition of quantity. The results of this review were used to define the treatment bundles described in the next section.

B. Treatment Bundles for Depression

To determine prices of treatment bundles for depression we use a data set consisting of insurance claims for four large self-insured employers that offered 25 health plans to 428,168 employees and their dependents. The data

were obtained from MEDSTAT, Inc., and contain information for the years 1991 through 1995. Information on drug claims, inpatient hospital treatment, outpatient visits, diagnoses, procedures, and demographic characteristics are reported. The health-insurance benefits offered to enrollees are generally quite generous relative to the market for private health insurance in the United States. The mental-health benefits are especially generous relative to typical private insurance. During the five years observed, there were important changes in the terms of mental-health insurance. While the majority of plans represent so-called managed indemnity plans (90–94 percent), the management of mental-health care changed for a substantial number of enrollees during the five years. Beginning on 1 January 1994, about 33 percent of the enrollees' mental-health coverage was "carved-out" to a specialty managed-care company. In January 1995, an additional 16 percent of enrollees had their mental-health benefits carved-out. These changes are expected to affect both the input prices and quantities of specific services delivered (e.g., visits).

In developing our treatment bundles we focus on the outpatient claims and the prescription-drug files. By focusing on outpatient treatment we reduce the number of observed severe cases of depression. Each outpatient and drug claim can accommodate two ICD-9 diagnostic codes. The point of departure was to identify cases of major depression. ICD-9 codes 296.2 (major depressive disorder, single episode) and 296.3 (major depressive disorder, recurrent episode) were used to define depression. Using the diagnostic information and dates contained in the claims, we construct episodes of treatment. In the case of prescription drugs, we consider the number of days of treatment provided by the prescription as the time period for which an individual received care. We follow previous research in identifying episodes of treatment as ending when no treatment is received for a period of time (Larry Kessler, 1980).

In defining our episodes of care we use an eight-week period without treatment to separate treatment episodes. Applying these criteria, we defined 20,603 episodes of care for the five years, 1991 through 1995. Censoring

of episodes occurred at both the beginning of 1991 and at the end of 1995. We do not consider the censored cases and therefore confine our attention to the 13,324 uncensored episodes. In order to limit the sample to less severe forms of major depression, we eliminated individuals with episodes involving inpatient hospital treatment at any time during the five years. This reduced the number of episodes to 10,368. Using information on procedures (e.g., type of visit) described by the Current Product Terminology (CPT) codes, we classify the composition of treatment that occurred within a treatment episode. Drug treatment is based on the National Drug Codes (NDC's) reported on the claim.

For this initial analysis we only consider "pure" treatments. That is, we only consider episodes of care that correspond directly to treatments tested in the clinical-trials literature. In this way we can directly link the "price" of an episode of a well-defined treatment to the price of other therapeutically similar treatments.

We identify nine major classes of treatment that have been shown to be efficacious in the treatment of depression: (i) psychotherapy alone, 6–15 visits; (ii) short-term TCA treatment alone or with medical management, 30–180 days; (iii) short-term SSRI treatment alone or with medical management, 30–180 days; (iv) short-term TCA treatment of 30–180 days with some psychotherapy; (v) short-term SSRI treatment (30–180 days) with some psychotherapy. The four remaining treatments are identical to the last four above, except for the provision of anxiolytic medication (Frank et al., 1998).

As previously mentioned, when claims data were converted to uncensored episodes of major depression, 10,368 episodes of depression were identified. Based on the definitions noted above, the number of episodes treated with each of the nine pure protocol treatments was calculated. It is notable that of the 10,368 episodes identified, a substantial share of treatments do not resemble any protocol treatment. For example, 1,818 episodes (47 percent of the 3,900) treated with psychotherapy alone consisted of a single visit. In addition, 1,672 or 16 percent of all episodes received neither psychotherapy nor an antidepressant drug. The

TABLE 1—AVERAGE COSTS OF TREATMENT IN 1991
(WITH 1995 COSTS IN PARENTHESES)

Treatment	Number of psychotherapy visits	Drug regimen (days)	N	Supply price (\$)	Demand price (\$)
Short-term psychotherapy	5–15	0	78 (197)	924 (646)	151 (95)
Short-term TCA	0	30–180	18 (8)	267 (117)	25 (39)
Short-term SSRI	0	30–180	33 (66)	254 (214)	11 (21)
Short-term psychotherapy and TCA	1–15	30–180	25 (13)	791 (391)	124 (59)
Short-term psychotherapy and SSRI	1–15	30–180	41 (128)	762 (582)	103 (90)

result was that the number of episodes receiving guideline standards of care appears to be relatively small (15–25 percent). The interpretation of the observed patterns of care is complex. For instance, in the case of single-visit episodes, those visits may have taken place for the purposes of “ruling out” major depression as the relevant condition to be treated in favor of a somatic condition or another mental disorder. In this case, the visit should not be viewed as “inappropriate care,” but as an appropriate assessment. The implication of this is that distinguishing between treatment and assessment is difficult in claims data, and so as much as 45 percent of treatment may lie close to the production frontier. In order to improve the precision of the estimated mean “prices” of treatment bundles, we aggregated several closely related bundles (e.g., short-term SSRI treatment with an anxiolytic drug and short-term SSRI treatment alone). The result was five treatment bundles used in the analyses reported here.

II. Prices of Treatment Bundles

In Table 1 we report the average supply (producer price index [PPI]) and demand (CPI) “prices” of each of the five treatment bundles in 1991 and 1995. The PPI in Table 1 measures the supply price, which includes both the payment made by the health plan and the patient’s out-of-pocket payments (OOP’s). The CPI reports the OOP’s or con-

sumer demand-price components for the same bundles (i.e., the co-payments or deductible contribution made by the patient/consumer).

Table 1 reveals rather dramatic differences in the supply “price” of treatment bundles for depression. This is even the case for therapeutically similar bundles. For example, short-term psychotherapy alone (5–15 visits) has an estimated price of about \$924 during the 1991 base year. Short-term TCA treatment alone (30–180 days) in contrast is priced at about \$267 if assessment and medical management costs are included; and SSRI alone (30–180 days) is slightly lower at \$254 per episode. Table 1 also shows even greater variation in the consumer demand price across treatments. For example, the demand price for the psychotherapy alone bundle is \$151, while TCA and SSRI treatments alone are about \$25 and \$11. From the patient’s vantage, therefore, the required OOP percentage is highest for psychotherapy alone and lowest for SSRI alone.

The most common form of mixed treatment is a combination of at least one psychotherapy visit along with a 30–180-day protocol level of treatment with an SSRI. In that case the supply price is \$762, and the demand price \$103 (14 percent of supply price). Psychotherapy with TCA is estimated to have a supply price of \$791 and a demand price of \$124 (16 percent of supply price). Thus the relative supply prices are quite comparable, largely due to the extra monitoring associated with TCA’s. Table 1 reports the supply and demand prices for 1995 in parentheses in the final two columns of the table. Nominal supply prices for all the five treatment bundles fell over the five years. In some cases the price decreases were substantial, such as the 30-percent fall in the price of psychotherapy alone. This was due primarily to a decrease in the price of a psychotherapy visit as opposed to a reduction in visits within the acceptable range (there were 8.05 visits in 1991 and 7.7 visits in 1995).

III. Construction of Price Indexes

Alternative formulas for constructing price indexes correspond to differing assumptions on the extent of substitutability among the treatment bundles. We examine four index formulations. One possible approach is to assume

that, in spite of their therapeutic similarity, the treatment bundles are completely nonsubstitutable, with idiosyncratic patients expected to respond only to one form of treatment. The Laspeyres fixed-weight base-period quantity index is implied by this zero-substitutability assumption. The fixed-weight Paasche index has this same property. For the fixed-weight Laspeyres index we use 1991 weights, whereas in the fixed-weight Paasche index we employed 1995 weights.

Two other alternatives involve less extreme assumptions. If one assumes that the elasticity of substitution between treatment bundles is unity, then one can construct the Cobb-Douglas index where expenditure share weights are computed as the mean expenditure shares for each of the five bundles over the 1991–1995 time period. Finally, as W. Erwin Diewert (1976) has shown, one can compute a Tornqvist discrete approximation to the continuous Divisia index that makes no a priori assumption about the elasticity of substitution among the five treatment bundles. Using the discrete Divisia index is consistent with the recommendations of the CPI Commission, whereas use of the other index numbers is not.

To this point we have not specified precisely what prices and quantities one would employ in these index-number calculations. In the case of the Laspeyres index, we use mean treatment bundle “prices” for each of the five aggregate bundles (built up from the nine efficacious bundles identified in the data). We follow the same approach for the other indexes as well. Note also that while the above discussion has focused on the construction of a PPI (supply price), the construction of a CPI (demand price) proceeds in an analogous manner.

IV. Results of Price Indexes, 1991–1995

Table 2 reports the results of constructing the PPI and CPI versions of the price indexes discussed above for five treatment bundles. The results for Laspeyres, Paasche, Cobb-Douglas, and Divisia indexes reported in Table 2 offer a consistent view of price movements for acute-phase treatment of major depression. Specifically, all four indexes indicate that there were both supply and demand “price” reductions of between 28 percent and

TABLE 2—PRICE INDEXES FOR TREATMENT OF DEPRESSION

A. PPI					
Year	Laspeyres	Paasche	Cobb-Douglas	Divisia	Psychotherapeutics, BLS
1991	100	100	100	100	100
1992	98.4	98.3	99.8	98.9	107.6
1993	86.7	88.9	87.4	85.6	113.4
1994	79.2	82.1	81.8	81.9	116.2
1995	68.4	71.9	70.6	71.2	120.4

B. CPI					
Year	Laspeyres	Paasche	Cobb-Douglas	Divisia	MCPI, BLS
1991	100	100	100	100	100
1992	91	87.6	94.3	91.2	107.5
1993	84	84.5	87.6	83.8	114.6
1994	80	80.9	85.3	82.3	120.5
1995	70	71.9	77.8	75.7	126.6

Note: Indexes normalized to 1991 = 100.

32 percent over the 1991–1995 period. The top panel of Table 2 shows that all of the PPI’s calculated experienced comparable falls to index values of between 68.4 and 71.9.

Among the dramatic changes occurring in the market for mental-health services, we expect to observe substitution across treatment bundles over time, from more psychotherapy-intensive and expensive care to greater utilization of lower-cost drugs and less psychotherapy-intensive care. One important trend is the shift toward treatments that make use of SSRI drugs as inputs. This is evidenced by the increasing shares of mixed treatments using SSRI drugs and the general growth in use of SSRI drugs alone. The diminishing role of psychotherapy alone is reflected by a declining share of sales. For example, from 1991 to 1994, the psychotherapy-alone bundle quantity share fell from 40 percent to 21 percent, although it rebounded to 47.8 percent in 1995. Finally, treatments that use TCA drugs have declined as a share of all treatments for depression. The PPI’s for depression can be compared to the Bureau of Labor Statistics’ (BLS) PPI for psychotherapeutic drugs reported in the right-most column of the upper panel of Table 2. The BLS index increased by 20 percent during the 1991–1995 period. The

average yearly differential in price changes is therefore on the order of 15 percent.

The lower panel of Table 2 reports the demand price or CPI results. As in the case of the PPI, the demand prices for acute-phase treatment of depression have moved downward during the 1991–1995 period. The price reductions differed somewhat across indexes. The fixed-weight Laspeyres index fell 30 percent compared to reductions of 28 percent, 22 percent, and 24 percent for the fixed-weight Paasche, Cobb-Douglas, and Divisia indexes, respectively. These substantial declines in the CPI for depression can be contrasted with the BLS medical CPI (or MCPI), which is reported in the last column of the lower panel of Table 2. The MCPI increased nearly 27 percent during the years 1991–1995. This appears to be due to (i) reliance on list prices as opposed to transaction prices (David Dranove et al., 1991) and (ii) use of fixed weights applied to the inputs (e.g., drugs, visits, etc.). The difference in yearly estimates of price changes is also about 15 percent. Both the supply- and demand-price growth differentials, from the BLS indexes, are about triple those reported by David Cutler et al. (1998) for heart-attack treatment.

The analyses of productivity in the mental-health sector and projected growth in spending typically relies on the MCPI and its components to project growth and composition of spending. An important implication of our results is that use of standard indexes may result in mistaking quantity changes for price changes. Therefore analyses of the composition of spending will result in a significant underestimate of growth in the quantity of effective care delivered and may also incorrectly project total spending into the future.

REFERENCES

- American Psychiatric Association.** "Practice Guidelines for Major Depressive Disorder in Adults." *American Journal of Psychiatry*, April 1993, 150(4), pp. 1–26.
- . *Diagnostic and statistical manual of mental disorders*, 4th Ed. Washington, DC: American Psychiatric Association Press, 1994.
- Busch, Susan; Frank, Richard G. and Berndt, Ernst R.** "Effectiveness, Efficacy, and Price Indexes for Depression: A Review of the Literature." Unpublished manuscript, Harvard Medical School, Department of Health Care Policy, 1996.
- CPI Commission.** *Final report to the Finance Committee from the Advisory Commission to Study the Consumer Price Index*. Washington, DC: Senate Finance Committee, 4 December 1996.
- Cutler, David M.; McClellan, Mark and Newhouse, Joseph P.** "The Costs and Benefits of Intensive Treatment for Cardiovascular Disease," in Jack E. Triplett, ed., *Measuring the prices of medical treatments*. Washington, DC: Brookings Institution Press, 1998 (forthcoming).
- Diewert, W. Erwin.** "Exact and Superlative Index Numbers." *Journal of Econometrics*, May 1976, 4(2), pp. 115–46.
- Dranove, David; Shanley, Mark and White, William D.** "How Fast Are Hospital Prices Really Rising?" *Medical Care*, August 1991, 29(8), pp. 690–96.
- Frank, Richard G.; Berndt, Ernst R. and Busch, Susan H.** "Price Indexes for Treatment of Depression," in Jack E. Triplett, ed., *Measuring the prices of medical treatments*. Washington, DC: Brookings Institution Press, 1998 (forthcoming).
- Griliches, Zvi.** "Quality Change and Index Numbers: A Critique." *Monthly Labor Review*, May 1962, 85(5), pp. 532–44.
- Kessler, Larry G.** "Episodes of Psychiatric Utilization." *Medical Care*, August 1980, 8(8), pp. 1219–27.
- Kessler, Ronald C.; McGonagle, Katherine A.; Zhao, Shanyang; Nelson, Christopher B.; Hughes, Michael; Eshlerman, Suzann; Wittchen, Hans-Ulrich and Kendler, Kenneth S.** "Lifetime and Twelve-Month Prevalence of DSM-III-R Psychiatric Disorders in the United States: Results from the National Comorbidity Survey." *Archives of General Psychiatry*, January 1994, 51(1), pp. 8–19.
- Scitovsky, Anne A.** "Changes in the Costs of Treatment of Selected Illnesses, 1951–65." *American Economic Review*, December 1967, 57(5), pp. 1182–95.
- Triplett, Jack E.** "Accounting for Health Care: Integrating Price Index and Cost Effectiveness Research," in Jack E. Triplett, ed., *Measuring the prices of medical treatments*. Washington, DC: Brookings Institution Press, 1998 (forthcoming).

Public Funds, Private Funds, and Medical Innovation: How Managed Care Affects Public Funds for Clinical Research

By JUDITH K. HELLERSTEIN*

Health research is produced by a variety of sectors of the economy: private for-profit industry, the government, and the nonprofit sector (primarily academic medical centers). These sectors are interdependent across many dimensions—even the funding for these sectors is not independent. In academic medical centers (AMC's), for example, which are the focus of this paper, a large amount of funding for research comes from the government, primarily through grants made by the National Institutes of Health (NIH), and from private for-profit industry like pharmaceutical firms. In addition, a substantial proportion of the research monies of AMC's has always come from the general revenues of the AMC's themselves, revenues that are raised primarily through the clinical treatment of patients in these institutions' faculty-practice plans or hospitals. In 1992–1993, for example, \$816 million (10 percent) of the faculty-practice-plan revenues of AMC's were used for research, an amount equal to 21 percent of the total NIH funding to these institutions in the same period (Robert F. Jones and Susan C. Sanderson, 1996). These general revenues, however, are being squeezed by recent changes in the structure of health insurance in the United States, specifically the rapid growth of managed care.

In this paper, I document the extent of government funding of AMC's for clinical research¹ over the last decade and consider how

managed care has affected clinical research through its effect on the ability of AMC's to raise government funds for research.

I. The Role of Managed Care and Government Funding of Clinical Research

Government funding for health research in AMC's can be thought of as an intermediate input into the production of medical innovations. While it might seem at first that government funding to AMC's should be a substitute for all other sources of funding, this is not at all clear. First, the NIH tries to fund research that would not have commercial applications, so research projects done with government and for-profit funds are often not substitutable. Moreover, it may be the case that government funding and general AMC revenues from hospitals and faculty-practice plans have been complements in production. Even when AMC's receive government funding for a research project, they must be able to give their faculty and staff time and other resources (such as infrastructure) that often cannot be accounted for in government grants and that therefore must come from general revenues.

When it comes to clinical research, AMC's must give participating physicians time to conduct the research, time that would otherwise be allocated to the revenue-raising activity of treating patients. To the extent that the growth of managed care may be reducing the general revenues of AMC's, managed care also may be adversely affecting the production of medical innovation by reducing the ability of

* Department of Economics, University of Maryland, College Park, MD 20742, and NBER. I thank Nancy Miller, James Schuttinga, and Judith Vaitukaitis for helpful discussions and for help in obtaining the NIH data, and Bob Moore for useful comments and for providing the NIH data. I also thank Kim Bayard and Lori Melichar for research assistance.

¹ NIH does not separately classify grants as being for clinical research. Instead, NIH classifies research grants on human subjects separately, since human-subjects grants must go through a special oversight panel before approval.

The data used in this paper technically refer to human-subjects grants, although I use the terms "human subjects" and "clinical" interchangeably since most human-subjects grants are grants for clinical research. The exceptions are grants for studies on human tissue and blood.

AMC's to conduct research with government funding.

Another way in which managed care may affect clinical research in AMC's is directly through the flow of patients to clinical research. When a patient is enrolled in a clinical research trial, the researcher is usually assumed to be responsible for costs incurred directly as a result of the research, but it is sometimes hard to properly categorize costs. Managed-care providers may therefore simply refuse to allow their patients to participate in clinical trials. Researchers are concerned about this issue, although there is little evidence that restrictions on patient participation are widespread (Robert E. Mechanic and Allen Dobson, 1996).

There have been a few case studies and general discussions of the impact of managed care on AMC's, including Mechanic and Dobson (1996), which looks explicitly at the direct effects on clinical research.² The only published study to date using national data to examine the effects of managed care on clinical research is Ernest Moy et al. (1997). This study graphs trends in various aspects of grants by NIH to AMC's, where AMC's are broken out as being in areas defined as having "high," "low," or "medium" managed-care penetration. The authors find that the growth rate of research grants by NIH has been slower in high-managed-care areas than in low-managed-care areas. The authors conduct no statistical tests of the differences between these two rates, nor do they control for other potentially confounding factors.

Although I provide no direct evidence in this paper on the importance of public research on medical innovation, it is worth emphasizing that there is a small but convincing body of evidence that public research has large positive spillovers to private innovation, both in medicine and in other scientific fields. Papers such as Adam Jaffe (1989), David Dranove and Michael R. Ward (1995), Francis Narin et al. (1997), and Andrew A. Toole (1997) all find that public research leads to higher lev-

els of private innovation. One caveat to this research, of course, is that it is difficult to measure correctly the impact of public innovations on overall consumer welfare.

II. The Data

The funding support for clinical research provided by NIH to AMC's over the last decade has been growing. NIH has provided me with data from their administrative data base on the numbers of grants and dollar awards of grants for research on human subjects made to each medical school in the United States in the years 1984–1995. From 1986–1994, the period considered in the data analysis below, the average annual growth rate of funding for human subjects in this period was 9.3 percent in nominal terms.

I have combined the NIH data on human-subjects grants made to each medical school in the United States with data that reflect the underlying demand for medical care in the markets in which these medical schools operate. The definition of a market is always subject to debate. For medical schools, it makes sense to define a market reasonably broadly since medical schools may draw from patient populations that come from farther away than usual. I therefore define the relevant market of a medical school as a consolidated metropolitan statistical area (CMSA) for medical schools in CMSA's or simply as a metropolitan statistical area (MSA), for medical schools located in smaller areas that qualify as MSA's but not as parts of CMSA's. Rural medical schools are eliminated from the sample, as are medical schools in Puerto Rico and the Uniformed Armed Services Medical School.

For each market assigned to a medical school, I use data from the 1997 Area Resource File (ARF) to calculate the levels and annual growth rates of variables that may be related to the demand for health care in that market. I calculate average annual levels and growth rates of (i) per capita HMO enrollment; (ii) per capita income; (iii) the percentage of the population over age 65;³ (iv)

² Other references include David Blumenthal and Gregg S. Meyer (1993), Marsha R. Gold (1996), and Lana R. Skirboll (1997).

³ The ARF only contains population estimates for the Medicare population for years starting in 1990. The data

population; (v) per capita income; and (vi) inpatient days in short-term general hospitals. All of these variables may affect the demand for the services of AMC's and therefore may affect AMC revenues. HMO's in the ARF include traditional HMO's, Independent Practice Associations (IPA's), and network and mixed HMO's. Unfortunately, the ARF HMO data, which are derived from various Interstudy surveys of HMO's, report enrollment in HMO's whose offices are located in a given geographical area, rather than actual HMO enrollment of the population in that area. Because I define a market reasonably broadly, this problem should not be too severe.⁴

The ARF also contains a large amount of other data on aspects of medical care in a given geographic area, such as the number of physicians in the area and the number of hospital beds. I experimented with including many of these variables in the regressions (both in levels and growth rates), but since they were never significant and since the rest of the qualitative results were unchanged by their inclusion, I do not report results using them. I restrict my analysis to the period 1986–1994, when I can get data on all the variables, and to AMC's that obtain NIH grants in each year of this period. Balancing the panel in this way does not change the qualitative results, as only a few small AMC's drop out. Summary statistics on the data appear in Table 1.

Grants to medical schools have cross-sectional variation (across AMC's) that dwarfs the variation in growth rates. In order to eliminate the fixed differences in grants across AMC's, I estimate only "differenced" regressions where I consider the impact of managed care on the *growth* of NIH grants to a medical school. I define the growth of grants to an AMC over the period 1986–1994 as the average annual growth rate in grants (averaged over all years of data).

It is not clear what aspect of managed-care penetration should most affect the growth rate

TABLE 1—SUMMARY STATISTICS FOR ACADEMIC MEDICAL CENTERS, 1986–1994

Variable	Mean	Standard deviation
NIH clinical research award dollars (in thousands)	16,128	17,059
Δ NIH clinical research award dollars	0.15	0.16
HMO enrollment	0.20	0.12
Δ HMO enrollment	0.13	0.11
Inpatient days (in thousands)	5,304	7,224
Δ inpatient days	-0.02	0.01
Population age 65+	0.12	0.02
Δ population age 65+	0.01	0.01
Income	19,815	2,764
Δ income	0.05	0.01
Population (in thousands)	4,990	5,996
Δ population	0.01	0.01

Notes: A unit of observation is an academic medical center (AMC). If an AMC is in a consolidated metropolitan statistical area (CMSA), all variables are measured at the CMSA level; otherwise, they are measured at the level of a metropolitan statistical area (MSA). For each observation, levels (growth rates) are measured as the average annual value (growth rate) for the period 1987–1994. All population and income variables are measured in per capita terms. There are 103 observations.

of NIH grants. If managed-care enrollment in a given year has a one-time impact on the ability of researchers to apply for and obtain NIH grants in that year, it is the (level) growth rate in HMO enrollment that should affect the (level) growth rate of NIH grants. However, it is likely that the level of managed-care enrollment in one year may have a dynamic impact on the ability of AMC's to get NIH grants for clinical research in future years (because the research process is dynamic, because faculty quality is serially correlated, etc.). In this case, the level of managed-care enrollment affects the growth rate of NIH clinical-research grants to medical schools.

The same argument can be made for other market demand factors that may affect clinical revenues of medical schools. In the empirical analysis, I therefore consider both the levels and growth rates of most of these variables and include them in separate regressions of the growth rate of awards on HMO enrollment. Controlling for these variables in an analysis of the effect of managed care on clinical research by AMC's is important since these fac-

I use to construct the levels and growth rates of the percentage of the population on Medicare are therefore from only this period.

⁴ For details on the definition of an HMO, and details on the rest of the ARF, see Bureau of Health Professions (1997).

TABLE 2—REGRESSION RESULTS (DEPENDENT VARIABLE—AVERAGE ANNUAL RATE OF GROWTH OF NIH CLINICAL-RESEARCH AWARD DOLLARS)

Variable	(i)	(ii)
Average level of HMO enrollment	-0.38 (0.14)	—
Δ HMO enrollment	—	0.07 (0.24)
Δ inpatient days	-0.68 (0.99)	-1.12 (0.74)
Δ population age 65+	-3.28 (3.26)	0.42 (3.27)
Δ income	-3.01 (3.09)	1.15 (3.75)
Δ population	-0.33 (1.86)	0.83 (0.20)
Constant	0.39 (0.21)	0.05 (0.19)
R^2 :	0.085	0.02

Notes: Standard errors (in parentheses) have been adjusted for correlation across AMC's within a CMSA or MSA. See notes to Table 1.

tors may be correlated with HMO enrollment in a geographic area.

III. Estimation Results

The main regression results are found in Table 2. In column (i), I report results from a regression of the average annual growth rate of NIH clinical-research award dollars to an AMC for 1986–1994 on the average level of HMO enrollment in the market area (as defined above) in which the AMC is located. I also control for the average growth rates of inpatient days, the fraction of the population on Medicare, per capita income, and the population itself. The point estimate of the coefficient on the HMO enrollment variable is -0.38 with a standard error of 0.14, which is statistically significant at the 5-percent level. None of the other coefficients in column (i) of Table 2 is anywhere close to being statistically significant. The R^2 of the regression is 0.09; this is not unexpectedly low given that the dependent variable measures growth rates. In an unreported regression, I replace the variables measured in growth rates in column (i) with the average levels of each of these variables over the sample period. The results are essen-

tially unchanged from the specification in column 1.

The economic magnitude of the statistically significant HMO effect is reasonably large. The point estimate on the HMO enrollment variable in column (i) implies that a one-percentage-point increase in HMO enrollment will lead to a 0.004-percentage-point reduction in the annual growth rate of NIH clinical research awards to a medical schools. Considering the mean annual growth rate in award dollars over this period, 0.15, this translates into a 2.5-percent reduction per year in the growth rate of awards.

Column (ii) of Table 2 reports results from a regression where I use as a measure of HMO enrollment the average annual *growth rate* of HMO enrollment. The point estimate on the growth rate of HMO enrollment is 0.07 and is not significantly different from zero. Once again, none of the other coefficients is statistically significant. This result, combined with the regression results in column (i) could be interpreted in two distinct ways. First, as I suggest above, this could imply that the HMO enrollment effect on clinical research does have a more dynamic structure, where it is the average HMO enrollment in an area that affects the growth rate of clinical research. The second interpretation is that average HMO enrollment in an area is a proxy for some other omitted geographic variable that is correlated with HMO enrollment and affects the growth rate in clinical research. It is therefore important to interpret the results with caution, although the fact that the other demand variables in the regression are insignificant lends some support for the idea that it is HMO enrollment itself that is driving the growth rates of clinical research.

IV. Conclusion

This paper provides empirical support for the concern of the medical community that the growth of managed care is adversely affecting the ability of AMC's to obtain NIH funding for clinical research. The results suggest that AMC's in areas with high managed-care penetration over the last decade have had lower rates of growth of NIH clinical research awards than other AMC's, even controlling for

other factors that may also affect the demand for medical care in these areas.

There are a few caveats. First, HMO enrollment is presumably not random, so the HMO effect in this paper could be a proxy for some other omitted variable. Second, the estimates in this paper are derived from "partial-equilibrium" regressions. There have been a few noted cases where prominent medical researchers have left AMC's in high-managed-care areas for AMC's in low-managed-care areas (see e.g., Alex Pham, 1996). In my estimates, this would imply a negative correlation between managed care and the growth of clinical research, when in fact this may simply reflect a shifting of resources among AMC's. Whether this shift is optimal, of course, is unknown.

There are many directions for future research. In the empirical work, I do not include any AMC-specific control variables in the regressions to help control for factors affecting specific AMC's. More importantly, my results on funding growth rates say nothing about other dimensions along which AMC's (or NIH) may be adjusting as a result of managed care. Establishing a direct link between the growth of managed care and the types of innovations being developed by the public sector should be a key topic for future research.

REFERENCES

- Blumenthal, David and Meyer, Gregg S. "The Future of the Academic Medical Center Under Health Care Reform." *New England Journal of Medicine*, December 1993, 329(24), pp. 1812-14.
- Bureau of Health Professions, Office of Research and Planning. *User documentation for the Area Resource File (ARF)*. Fairfax, VA: Quality Resource Systems, Inc., February 1997.
- Dranove, David and Ward, Michael R. "The Vertical Chain of Research and Development in the Pharmaceutical Industry." *Economic Inquiry*, January 1995, 33(1), pp. 70-87.
- Gold, Marsha R. "Effects of the Growth of Managed Care on Academic Medical Centers and Graduate Medical Education." *Academic Medicine*, August 1996, 71(8), pp. 828-38.
- Jaffe, Adam. "Real Effects of Academic Research." *American Economic Review*, December 1989, 79(5), pp. 957-70.
- Jones, Robert F. and Sanderson, Susan C. "Clinical Revenues Used to Support the Academic Mission of Medical Schools, 1992-1993." *Academic Medicine*, March 1996, 71(3), pp. 300-7.
- Mechanic, Robert E. and Dobson, Allen. "The Impact of Managed Care on Clinical Research: A Preliminary Investigation." *Health Affairs*, Fall 1996, 15(3), pp. 72-89.
- Moy, Ernest; Mazzaschi, Anthony J.; Levin, Rebecca J.; Blake, David A. and Griner, Paul F. "Relationship Between National Institutes of Health Research Awards to US Medical Schools and Managed Care Market Penetration." *Journal of the American Medical Association*, July 1997, 278(3), pp. 217-21.
- Narin, Francis; Hamilton, Kimberly S. and Olivastro, Dominic. "The Increasing Linkage Between U.S. Technology and Public Science." *Research Policy*, December 1997, 26(3), pp. 317-30.
- Pham, Alex. "Medical Brain Drain: Researchers Leaving Boston Say Managed Care Economics Are Drying Up Seed Money." *Boston Globe*, 13 July 1996, p. 65.
- Skirboll, Lania R. "The Impact of Managed Care on Research: The Changing Face of Medicine." *Academic Medicine*, September 1997, 72(9), pp. 778-79.
- Toole, Andrew A. "The Impact of Federally Funded Basic Research on Industrial Innovation: Evidence from the Pharmaceutical Industry." Mimeo, Laurits R. Christensen Associates, Madison, WI, 1997.

The Demand for Medical Care: What People Pay Does Matter

By MATTHEW J. EICHNER *

The movement toward managed care is likely the most important development in the U.S. health-care system over the past two decades. Managed-care providers have typically lowered prices to consumers beneath the coinsurance level under traditional fee-for-service coverage while installing a complicated structure to exercise administrative control over the amount and nature of care provided. Yet there is now widespread dissatisfaction with some of these cost-control mechanisms. The popular press is full of stories on the supposed excesses of these administrative-control systems, and in the recent election cycle, an initiative appeared on the ballot in California to outlaw certain practices. The recent financial difficulties of managed-care organizations in a number of markets have also focused attention on whether supply-side controls will prove able to contain costs in the long run.

An alternative approach to controlling costs relies on inducing or compelling individuals to bear a greater share of the marginal cost of treatment in the hope that they will consume only those services which they value at some level near to cost. Understanding the behavioral response of the insured to price incentives is essential to understanding the potential health and economic implications of such market-based schemes. Yet there is relatively little information in the empirical literature about how individuals respond to changes in the out-of-pocket cost of medical care. The dearth of information is reflected in the fact

that recent papers on catastrophic health insurance and medical savings accounts continue to cite empirical estimates of the response to the out-of-pocket price of care produced from the Rand health-insurance experiment two decades ago.

The Rand experiment used specialized sample-design and data-collection techniques to produce credible estimates of plan effects with attractively simple statistical techniques. In this paper, I describe statistical methods to allow the computation of credible estimates from less specialized data. I produce estimates of the behavioral response to changes in the out-of-pocket cost of care using only the claims data that are routinely collected by employers and insurers.

My methods rely on two characteristics of employer-provided medical insurance in the United States. First, plans offer a multilevel schedule of reimbursement. In a typical plan, the first several hundred dollars of medical expenditures per year (the deductible) are not reimbursed and must be paid out-of-pocket. After this deductible is met, the insurance reimburses a fixed percentage of the cost of the medical care consumed, usually between 70 percent and 90 percent. Then, once a still higher level of expenditure is reached, the insurance plan pays all subsequent charges. Thus the fraction of costs paid out-of-pocket within a single plan varies between 1 and 0. Second, another feature of medical insurance is that whole families, and not single individuals, receive insurance coverage and are subject as a unit to the schedule of deductibles and copayments.

These characteristics of insurance allow estimation of a behavioral response to the price of medical care. My basic strategy involves recognition of the fact that expenditures by an employee's family reduce the price he or she must pay for medical care. Suppose, for example, that employees at a firm are enrolled in a plan with a \$500 deductible after which they pay only 20

[†] *Discussants:* Mark Pauly, University of Pennsylvania; Roger Feldman, University of Minnesota.

* Department of Economics and Finance, Columbia University Graduate School of Business, 612 Uris Hall, New York, NY 10027. This research relied on medical claims data which, for reasons of confidentiality, cannot be made available except by special arrangement with the provider of the data.

percent of the cost of care out-of-pocket. I compare the spending of employees who are equivalent in all observable characteristics except for the level of medical expenditures by other family members. Those whose dependents have accumulated over \$500 in expenditures during the year face a marginal cost of 20 percent, while those whose dependents have been healthy face a marginal cost of 100 percent. I interpret the difference in expenditures between the two sets of employees as a behavioral response to variation in the cost of medical care.

I. Firm Medical Claims Data

The data used in the following analysis are collected by firms in order to process the claims of employees for the reimbursement of medical costs. Each record represents a specific claim for a specific service on a specific date and indicates the identity of the individual receiving the service, the household to which he or she belongs, the plan under which the patient is covered, the diagnosis, the type of service, and the billed cost of care rendered. Basic demographic information, including the patient's age, gender, and location, is also included as part of each claim record. Claims data have the attractive property that individual plan enrollees have a motivation to report every claim. At the same time, the firm paying the bill has an incentive to make sure that only legitimate claims are filed.

The data used in this paper come from a Fortune 500 firm employing 16,989 workers between the ages of 25 and 55 in five business units. These employees, along with their dependents, filed approximately 487,000 claims per year during the 1990–1992 period. The differences among the three plans are completely captured by the differences in deductibles, copayments, and stop-loss limits. Utilization-review procedures for certain high-cost treatments and “carve-outs” for mental health, substance abuse, eyeglasses, and prescription drugs are the same across the three plans. Further discussion of these data is found in Eichner (1997).

II. Identifying a Behavioral Response to the Price of Health Care

The methodological approach taken in this paper differs markedly from other efforts to

understand the behavioral response to price incentives, most notably the Rand health-insurance experiment. Rather than identify a response by comparing individuals across plans, this approach isolates the behavioral parameters of interest using variation in out-of-pocket cost within a plan with an annual deductible during a calendar year. Due to the structure of these plans, the price of care under insurance plans with annual deductibles can fall during the calendar year.

To use this variation, I must separate the change in price into two components. Some of the fall in price during a year is due to an employee's own expenditures. By spending on health care and satisfying deductibles and stop-loss limits, employees reduce their out-of-pocket cost for additional care. This variation creates an endogeneity problem, as spending by the employee on his or her own medical care and thus the price paid for additional care are surely correlated with his or her unobservable health status. However, there is also variation in cost that is not due to an employee's own expenditure but, because of the manner in which employer-provided insurance is structured, to the expenditures of an employee's family.

Consider two observationally identical families covered by an insurance plan with a \$250 annual deductible and a 20-percent copayment. On January 1, celebrating the New Year, a dependent in one family jumps from the upper tier of a bunk bed and requires a \$300 trip to the emergency room. Suddenly, one employee faces a marginal cost of care of \$0.20, while her seemingly identical colleague faces a marginal cost of \$1.00. Measuring the difference in behavior which results from such variation in cost is the topic for the remainder of this paper.

Obviously, not all expenditures are like a flight from the bunk bed in that they produce a change in the price of care faced by the employee that is independent of the employee's own medical history. Some expenditures of family members, like those for high-potency antibiotics when strep infections sweep a day-care center, signal medical events that affect all members of the family, including the employee. Other expenditures of family members, such as those for treatment of the

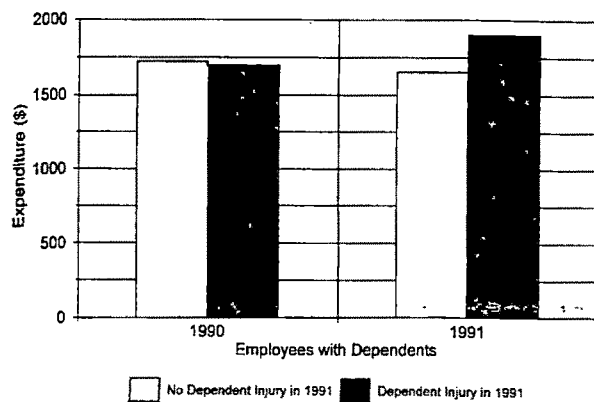


FIGURE 1. MEAN MEDICAL EXPENDITURE BY 1991 DEPENDENT INJURIES

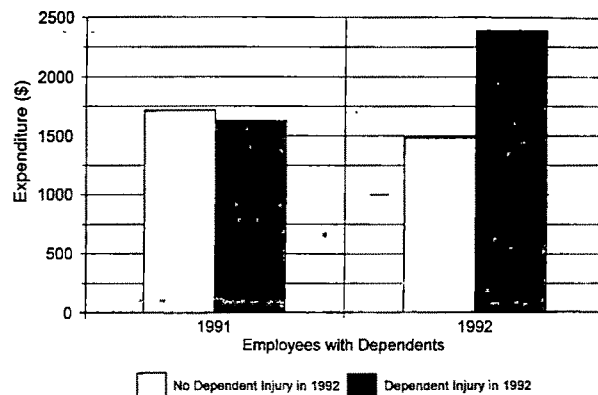


FIGURE 2. MEAN MEDICAL EXPENDITURES BY 1992 DEPENDENT INJURIES

common cold, signal an underlying propensity of the family to consume care which is likely both to lower the price faced by the employee and to cause her to visit the doctor frequently.

My solution to both of these potential problems is to use only the subset of medical expenditures resulting from injuries and poisonings. I do this in the hope that these claims of family members produce variation in the price of care faced by the employee that are unconnected with her own medical history. In other words, these are the claims resulting from flights from the bunk beds. To avoid the problem of events that affect multiple family members, I discard claims for injuries and poisoning for which multiple family members receive treatment at the same time.

Figure 1 depicts the identification strategy described above applied to firm employees with one or more dependents. The left two bars show mean employee expenditures for 1990. Expenditures are shown conditional on whether the employee's dependents incurred expenditures for injuries, accidents, or poisonings during the year 1991. The right bars show mean expenditures for 1991, again conditional on whether dependents incurred expenditures for injuries, accidents, or poisonings during that calendar year. The left two bars are essentially indistinguishable, suggesting that expenditures looked the same in 1990 whether or not dependents incurred injury and poisoning expenditures in 1991. The right two bars, however, are quite distinguishable, with those employees whose dependents incurred expenditures themselves spending more. I interpret

this difference between the right two bars as a behavioral response to the price of care which has been reduced by the exogenous expenditures of dependents satisfying the annual deductible. Figure 2 shows an analogous pattern, but for the years 1991 and 1992.

III. Estimating the Behavioral Response to Out-of-Pocket Price

I will apply the identification strategy outlined in the previous section to estimating the behavioral response to the out-of-pocket cost of medical care. Here I will take as the relevant cost the year-end out-of-pocket cost or, more precisely, the derivative of total out-of-pocket health-care cost with respect to the marginal dollar of care. Consider the employee who spends an additional dollar of medical care on December 31. The fraction of that dollar paid out-of-pocket represents this derivative. If the deductible is not yet satisfied, this fraction is 1; if the deductible is satisfied the fraction is the copayment percentage. One way to motivate this particular cost measure is to assume that individuals are endowed with perfect foresight regarding, if not the total amount of family medical consumption during the calendar year, at least whether this total will exceed the deductible and, if so, whether it will exceed the stop-loss limit. Alternative cost measures which treat the year-end price as uncertain are introduced in Eichner (1997).

As explained above, there is a simultaneity problem with this cost measure: bad health and the resulting medical costs lower the

out-of-pocket cost of care. At the same time, low out-of-pocket costs for care likely induce additional consumption of health care and higher medical costs. To separate the two directions of causality, I use an indicator for whether dependents incurred expenditures for injuries, accidents, or poisonings as an instrument for the price the employee faces at the end of the year. Since I am interested in using injury and poisoning events to break the correlation between spending by an employee and spending by his or her dependents, I will omit approximately 200 cases related to injuries and poisoning which affected multiple family members.

I estimate a basic relationship between expenditure and out-of-pocket cost using the minimum-distance method described in Whitney K. Newey (1987):

$$\begin{aligned} (1) \quad \log(\text{Exp}) \\ &= \alpha + \beta_1 \log(\text{Price}) + \mathbf{Z}\beta_2 \\ &\quad + [1(\text{Plan} = 1) \times \mathbf{Z}] \beta_3 \\ &\quad + [1(\text{Plan} = 3) \times \mathbf{Z}] \beta_4 + \mu \end{aligned}$$

where Exp is employee expenditure and Price is the end-of-year marginal cost. The vector \mathbf{Z} contains indicators for age, gender, and business unit, as well as an age \times gender interaction term. Variables included in \mathbf{Z} are interacted plan indicators.

Table 1 summarizes the estimates obtained and presents these along with the implied demand elasticities. In all cases, equation (1) was evaluated at the mean values of all regressors to produce the elasticity estimates. The estimate constructed using data from all employees for the year 1992 implies, for example, that a 1-percent increase in out-of-pocket cost produces a 0.62-percent fall in expenditure.

IV. Discussion of Results

The estimates shown in Table 1 are somewhat larger than other recent efforts to measure the response to the out-of-pocket price of medical care, although not of a different order of magnitude. The Rand health-insurance experi-

TABLE 1—SUMMARY OF PARAMETER ESTIMATES AND IMPLIED ELASTICITIES

Group	Parameter	1990	1991	1992
All employees	tobit	-2.0310 (0.0212)	-2.1631 (0.0214)	-2.4239 (0.0229)
	minimum distance	-1.0048 (0.0454)	-1.0327 (0.0450)	-0.8814 (0.0452)
	implied elasticity	-0.7359	-0.7452	-0.6151
Employees with one or more dependents	tobit	-2.2967 (0.0326)	-2.0010 (0.0311)	-1.8611 (0.0315)
	minimum distance	-1.0689 (0.0699)	-0.9740 (0.0667)	-0.8034 (0.0692)
	implied elasticity	-0.7851	-0.7149	-0.5671

Note: Numbers in parentheses are standard errors.

ment, which randomized families into different plans with differing cost-sharing arrangements, found an overall elasticity for medical care of about -0.22 (Joseph P. Newhouse, 1993). Eichner (1997) applies a similar methodological approach to that described in this paper, but using a more complicated price variable which takes into account the evolution of expectations concerning prices through the calendar year until December 31, when the year-end price is revealed. Here too, the estimated elasticities are smaller, ranging between -0.22 and -0.32.

I believe that the method described in this paper will prove useful in addressing a number of policy-relevant questions that are difficult to answer without the large sample sizes characteristic of administrative data. For example, this approach is well suited to investigating the behavioral response to price for different types of medical care. It is also possible to investigate how responses vary in different populations. Finally, using administrative data in this fashion will allow not only measurement of the movement in central tendency when the price of care changes, but also the effects of price at different points in the distribution. Given the extremely skewed nature of the medical expenditure distribution, understanding where in the expenditure distribution the effects of out-of-pocket cost are large is important in evaluating a number of options for reform of the U.S. health-insurance system.

REFERENCES

- Eichner, Matthew J. "Incentives, Price Expectations, and Medical Expenditures: An Analysis of Claims Under Employer-

Provided Health Insurance." Mimeo, Massachusetts Institute of Technology, 1997.

Newey, Whitney K. "Efficient Estimation of Limited Dependent Variable Models with Endogenous Explanatory Variables." *Jour-*

nal of Econometrics, November 1987, 36(3), pp. 231-50.

Newhouse, Joseph P. *Free for all: Lessons from the RAND health insurance experiment*. Cambridge, MA: Harvard University Press, 1993.

Adverse Selection and Adverse Retention

By DANIEL ALTMAN, DAVID M. CUTLER, AND RICHARD J. ZECKHAUSER *

Most employers providing health insurance offer a menu of plans and allow employees to choose the plan they prefer. As a result, the demographic mixes of insureds and, consequently, costs differ dramatically across plans. The average 60-year-old, for example, spends more than twice as much annually as the average 30-year-old. A single high-cost individual can incur costs equal to the total of several hundred low-cost insureds. Because of this variability, insurers are deeply concerned about how people choose their plans, as are employers and governments that finance and monitor them, and analysts who study them. We seek to understand who insures in which health-insurance plan, and why they do it. This information will enable us to correctly calculate cost differentials between plans, and to set premiums accordingly.

The mix of people in a plan in a particular year depends on who began there, and who moves in and out. Economists have long been fascinated by the movers. The tendency for sick (healthy) people to join plans offering rich (lean) services at high (low) cost is termed "adverse selection" (Michael Rothschild and Joseph Stiglitz, 1976; Charles Wilson, 1987). Most discussion of adverse selection, concentrating on the decision of whether to purchase insurance, focuses on high risks moving to very generous plans, with the low risks choosing to go without insurance coverage. But for mandatory or heavily subsidized insurance, such as employer-provided health care, everybody will be insured, and

thus other groups might be important as well: low-risk movers seeking lower prices, new enrollees, and people who stay put.

We focus in this paper on why premiums differ so much across health-insurance plans. We show that adverse selection is quantitatively important, but that it is more a result of low-risk people moving out of generous plans than of high-risk people moving into those plans. We then document the opposite of adverse selection, a concept we term "adverse retention." Adverse retention is the tendency for people who stay put to magnify cost differentials between plans, as they will if they differ in age and costs are more than linear with age. We show that adverse retention has about 60 percent as large an effect on health-plan premiums as does adverse selection.

I. The Setting

To analyze the factors accounting for differences in plan premiums, we obtained data on plan enrollment and utilization for people insured through the Group Insurance Commission (GIC) of Massachusetts. The GIC provides health insurance to state and local employees in Massachusetts. With roughly 133,000 employees and 245,000 total lives, it is among the largest insurance purchasers in New England (see Cutler and Zeckhauser [1998] for more description). We secured complete expenditure and plan membership records for GIC enrollees for the 30-month period from July 1993 through December 1995 (fiscal years 1994, 1995, and half of 1996).

The GIC offers three types of health-insurance plans. A traditional indemnity policy is the most generous, imposing few restrictions on utilization and only moderate cost-sharing. The second plan is a Preferred Provider Organization (PPO). Its enrollees are steered toward "network" providers, but the providers are still paid on a fee-for-service basis, and the restrictions on utilization are mild. Ten HMO's comprise the most stringent plans.

* Altman: Department of Economics, Harvard University, Cambridge, MA 02138; Cutler: Department of Economics, Harvard University, Cambridge, MA 02138, and National Bureau of Economic Research; Zeckhauser: Kennedy School of Government, Harvard University, Cambridge, MA 02138, and National Bureau of Economic Research. We are grateful to Charles Slavin and Roger Feldman for helpful discussions and to the National Institutes on Aging for research support. The data used in this paper are proprietary.

TABLE 1—PLAN PREMIUMS, ENROLLMENT, AND BENEFIT COSTS

Plan	Premium (\$)	Enrollment	Benefit cost (\$)	Adjusted benefit cost (\$)
Indemnity	2,670	66,253	2,176	1,908
PPO	1,631	24,032	1,115	1,202
HMO	1,686	117,501	1,233	1,320
Difference:				
Indemnity — HMO	984	—	943	588

Notes: Data are for fiscal year 1995. Adjusted benefit costs control for differences in the age and sex mix of each plan. Enrollment and benefit costs include only individuals under age 65.

HMO enrollees are required to use particular providers, who are paid on a salary or "capitation" (a single amount per patient per year) basis. The HMO's rely heavily on utilization review.

The first column of Table 1 shows average per capita premiums (employer plus employee payment) in fiscal year 1995. The indemnity policy is significantly more expensive (by about \$1,000 per person per year) than the other policies. The difference between the indemnity and HMO premiums has been substantial for at least the past decade; the PPO has been priced at about the level of the HMO's since just after it was introduced.¹ Despite its far greater cost, enrollment in the indemnity plan has remained relatively stable at about 32 percent of total nonelderly enrollees (the second column of the table). The bulk of the remainder enroll in an HMO; as a result, we focus on indemnity-HMO differences.

The third column shows average claims paid in the different plans (termed "benefit costs"). Benefit costs are much lower than premiums; part of the difference may arise because we are missing some claims information. But more importantly, benefit costs differ by virtually the same amount across plans as

plan premiums. If administrative expenses varied significantly across plans, premiums less benefit costs would vary in like amount. Table 1 suggests that administrative expenses do not vary significantly across plans.

A central question for insurance plan design is why the indemnity policy is so expensive relative to the PPO and HMO's. The two major possibilities are differences in population mix and management differences (e.g., more stringent utilization management or lower prices paid in the HMO's). In ongoing work, we are comparing the management of similar conditions across plans. Our focus in this paper is on the effect of population mix across plans.

To investigate the role of demographics in explaining premium differences, we form a measure of "adjusted" benefit costs. We take plan-specific spending by age and sex and then estimate benefit costs as if the age and sex mix in the GIC as a whole were enrolled in each plan. The difference between adjusted and unadjusted benefit costs indicates the importance of demographics in explaining premium differentials. A comparison of the last two columns shows that demographics are very important for cost differences; demographic differences explain 38 percent of the difference in plan costs ($[1 - \$588/\$943] \times 100$).²

II. Adverse Selection and Adverse Retention

How much of the cost differential between the indemnity plan and the HMO's is due to people switching plans in light of their health state (adverse selection), and how much is due to the fact that people within plans differ in mix and do not switch plans (adverse retention)? Table 2 provides evidence about forces affecting the enrollment mix. We show statistics on people in the fiscal years 1994–1995 period, and the fiscal years 1995–1996 period.

¹ The PPO was first offered in 1994. It was priced high that year because enrollment was uncertain. Favorable experience in 1994 led the plan sponsors to lower the premium. It has remained about the level of the HMO's since 1995.

² Similar conclusions about the importance of demographics in the GIC are obtained by Arlene Ash et al. (1997). In other groups, demographics are less important in plan premiums (Randall P. Ellis, 1989). Beyond demographics and management differences, the incidence of disease may vary across plans. In ongoing work, we are exploring this (see also Wei Yu et al., 1998).

TABLE 2—CHARACTERISTICS OF MOVERS AND STAYERS

Characteristic	FY 1994–FY 1995		FY 1995–FY 1996	
	Indemnity	HMO	Indemnity	HMO
Number of people				
Stayers	62,369	108,369	59,725	109,859
Movers (number in initial plan)	1,474	1,039	1,367	734
New enrollees	2,452	7,058	1,722	6,283
Average spending				
Stayers, first year	\$2,252	\$1,125	1,981	1,327
Movers, first year	<i>1,444*</i>	<i>1,651*</i>	1,381*	1,385
Stayers, second year	\$1,960	\$1,344	—	—
In-migrants, second year	<i>2,095</i>	<i>1,181</i>	—	—
New enrollees, second year	1,433*	1,042*	—	—

Notes: Numbers are adjusted for the age and sex mix of the GIC as a whole. Movers in the first year are in-migrants in the second year, as the italics and underlining indicate.

* Significantly different ($P < 0.05$) from the value for stayers.

The first rows report the number of people staying in their health plan, the number moving to new plans, and the number of new enrollees. About 2 percent of people who were in the indemnity plan moved to an HMO each year, and about 1 percent of HMO members moved in the reverse direction.³

The next rows show costs for these groups of people, adjusting for demographic differences across the groups.⁴ We present spending data for fiscal year (FY) 1994 and FY1995; because we only have claims for half of FY1996, we do not present data for that year. Adverse selection is clearly present; movers are far from representative of their plan. People leaving the indemnity plan for an HMO, for example, spent 30–36 percent less than people who remained in the indemnity plan in each year (for example, \$1,444 vs. \$2,252 in FY1994). This is true in both years of the data.⁵ Adverse selection is present among

³ The fact that plan mobility is so low, as it is in many other groups (Joachim Neipp and Zeckhauser, 1985), suggests that adverse retention might be important.

⁴ We also examined hospitalization probabilities for the different groups and found similar results.

⁵ Similar findings are true for Medicare. People who joined an HMO spent only 63 percent as much in the six months prior to HMO enrollment as people who remained in the fee-for-service system (Physician Payment Review Commission, 1996). But in the Medicare program, no data are available on spending while enrolled in an HMO.

TABLE 3—EFFECT OF ADVERSE SELECTION AND ADVERSE RETENTION ON BENEFIT COSTS

Effect of:	Indemnity	HMO	Net
Adverse selection			
Movers	\$16	–\$9	\$25
New enrollees	\$20	\$19	\$1
Adverse retention	\$23	\$9	\$14

Note: Data are adjusted for demographic differences across groups.

movers in the other direction as well; people who left an HMO for the indemnity plan after FY1994 used 47-percent more services in FY1994 than those remaining in an HMO the next year, although people who left the HMO after FY1995 were only about average compared to those who remained in an HMO in FY1996.

Once people change plans, however, their spending is only about average for enrollees in the new plan. People who move from an indemnity plan to an HMO in FY1995, for example, incur approximately average costs for HMO enrollees that year (\$1,181 vs. \$1,344); the same is true for people who join the indemnity plan.

To allow a rough calculation of the importance of adverse selection, we assume that there was no mobility between FY1994 and FY1995, and that the movers would have spent the same share relative to the stayers in FY1995 as they did in FY1994. As Table 3 shows, without mobility the indemnity plan's average costs would have been \$16 lower (about 1 percent). Spending in the HMO's would have been \$9 higher (about 1 percent). The relation of these two numbers is striking. Two-thirds of total adverse selection results from low-cost people moving out of the generous plan; only one-third results from high-cost people moving into the generous plan.

In addition to adverse selection from movers, there may also be adverse selection from joiners (new enrollees). To estimate this component of selection, we find the effect of the new enrollees on premiums. As Table 3 shows, new enrollees reduced costs in the indemnity policy by \$20 per person and costs in the HMO's by \$19 per person. While these

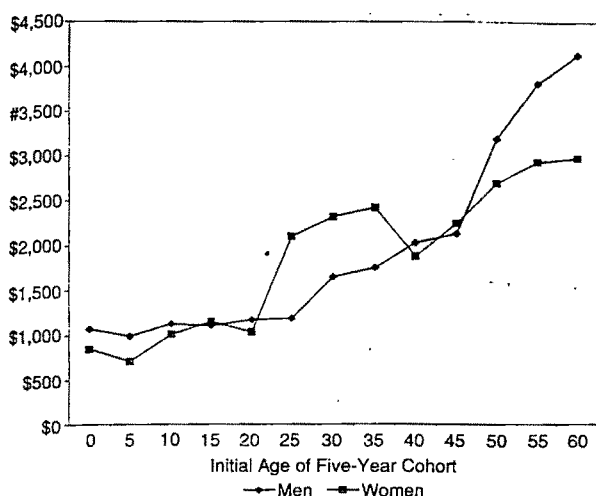


FIGURE 1. MEDICAL SPENDING BY AGE

effects are large, they are virtually the same for the two plans. Adverse selection results primarily from people switching plans, not from new enrollees disproportionately moving into certain plans.

We next explore the impact of adverse retention. Adverse retention will only affect plans differently if medical spending is non-linear in age. Figure 1 shows that this is the case; for men, in particular, costs increase quite rapidly between ages 55 and 64. Plans with people aging into or within this group will therefore experience above-average increases in medical spending. To evaluate the impact of adverse retention, we simulate the effect of the stayers on plan premiums. We take the FY1994 demographic distribution of stayers in each plan and age them by one year, holding constant age- and sex-specific spending by plan. As Table 3 shows, the increasing age of the stayers would raise the cost of the indemnity plan by \$23 (about 1 percent) and would raise the cost of the HMO's by \$9 (about 1 percent). The net effect of adverse retention in driving cost differentials is therefore about \$14. Thus, adverse retention is roughly 60 percent as important as adverse selection in driving up the costs of the indemnity plan.

In addition to adverse selection and adverse retention, demographic differences have a direct effect on plan spending. Based on demographic mix alone, if costs in the indemnity policy applied to all individuals, we estimate that those enrolled in HMO's would spend

\$390 less per person than would those with indemnity policies.⁶

III. Implications

Differences across plans in the mix of enrollees are a serious concern in health-insurance markets. These differences are large and have increased over time. Changes in the demographic mix result from adverse selection, but also significantly from adverse retention: the fact that few people change plans, and costs rise at an increasing rate with age. We estimate that adverse retention is roughly three-fifths as important as total adverse selection.

Adverse selection and adverse retention will vary in importance from group to group. While data for other groups are scarce, we suspect that the degree of adverse selection and retention will depend on three factors: the provisions and management ability of the different plans, the heterogeneity of the population, and the length of time for which the plans have been offered. The last aspect is particularly important if people do not switch plans frequently; adverse retention will drive up the costs of older plans relative to younger plans.

Biased enrollments in GIC plans have proved consequential. In most recent years, the indemnity plan has increased in cost relative to the HMO's and has lost enrollment. The GIC is required to offer an indemnity plan and, therefore, to fight the forces of adverse selection and retention.⁷ The GIC's primary strategy has been to cross-subsidize the indemnity plan, by paying 85 percent of plan premiums. Thus, individuals can join the indemnity plan by paying only 15 percent of the cost differential.⁸ The

⁶ This calculation differs from Table 1 because it is for stayers only, while Table 1 is for all enrollees.

⁷ It is widely believed that most state legislators (who have required the indemnity plan to be continued) are in the indemnity plan and thus would be concerned about its survival and its rate.

⁸ Many other employers have a policy of paying a fixed amount for health insurance, independent of the plan chosen. If employees had to pay the full incremental cost of the indemnity plan, adverse selection would be more severe (Roger Feldman et al., 1989; Cutler and Sarah Reber, 1998).

GIC has also actively managed the costs of the indemnity plan, including "carving out" mental-health benefits and prescription-drug benefits to lower-cost providers; bargaining with hospitals over laboratory and other ancillary services; and actively managing care for high-cost outpatient users.

Efficiency demands that the GIC, and more generally employers offering health plans, should charge insureds amounts intended to reflect true cost differentials across plans. Moreover, many would say equity is also served if those who want more services are required to pay for them. The GIC approach of neglecting risk mix and subsidizing cost differentials heavily employs two counterbalancing wrongs, which only miraculously could come out right. It also provides few incentives for plans to compete vigorously on the basis of price, since employees realize only 15 percent of the gains of any price reduction.

Economists typically prescribe implementing "risk adjustment" in such situations. This would involve a greater employer contribution per capita for plans that enroll more high-risk people and less to plans that enroll more low-risk people. If payments from the GIC were calibrated to offset risk differences, employees seeking higher-service plans would then be charged amounts reflecting only the management differences across plans for a standardized population mix. Risk adjustment can be prospective (e.g., paying more to plans with more heart-attack survivors), retrospective (e.g., subsidizing high-expense cases through a reinsurance system), or a mixture of the two. Risk adjustment is just beginning to be recognized, and its methods implemented. Designing and testing alternative risk-adjustment systems will be critical if we are to encourage choice and competition in health insurance.

REFERENCES

- Ash, Arlene; Ellis, Randall P.; Yu, Wei; Mackey, Elizabeth; Iezzoni, Lisa; Ayanian, John; Bates, David; Burstin, Helen; Byrne-Logan, Susan and Pope, Gregory. "Risk Adjusted Payment Models for the Non-elderly." Report to the Health Care Financing Administration, Washington, DC, September 1997.
- Cutler, David M. and Reber, Sarah. "Paying for Health Insurance: The Tradeoff Between Competition and Adverse Selection." *Quarterly Journal of Economics*, 1998 (forthcoming).
- Cutler, David M. and Zeckhauser, Richard J. "Adverse Selection in Health Insurance," in Alan Garber, ed., *Frontiers in health policy*. Cambridge, MA: MIT Press, 1998 (forthcoming).
- Ellis, Randall P. "Employee Choice of Health Insurance." *Review of Economics and Statistics*, May 1989, 71(2), pp. 215-23.
- Feldman, Roger; Finch, Michael; Dowd, Bryan and Cassou, Steven. "The Demand for Employment-Related Health Insurance Plans." *Journal of Human Resources*, Winter 1989, 24(1), pp. 115-42.
- Neipp, Joachim, and Zeckhauser, Richard J. "Persistence in the Choice of Health Plans," in R. M. Scheffler and L. F. Rossiter, eds., *Advances in health economics and health services research*, Vol. 6. Greenwich, CT: JAI Press, 1985, pp. 47-74.
- Physician Payment Review Commission. *Annual Report to Congress*. Washington, DC: U.S. Government Printing Office, 1996.
- Rothschild, Michael and Stiglitz, Joseph E. "Equilibrium in Competitive Insurance Markets." *Quarterly Journal of Economics*, November 1976, 90(4), pp. 630-49.
- Wilson, Charles. "Adverse Selection," in John Eatwell, Murray Milgate, and Peter Newman, eds., *The new Palgrave dictionary of economics*. London: Macmillan, 1987, pp. 32-34.
- Yu, Wei; Ellis, Randall P. and Ash, Arlene. "Using the Diagnostic Cost Groups (DCGs) To Measure Risk Selection in the Massachusetts State Employee Health Insurance Program." Mimeo, Boston University, 1998.

Payment Heterogeneity, Physician Practice, and Access to Care

By SHERRY GLIED*

Today's health-care marketplace is characterized by a historically remarkable array of reimbursement systems, practice rules, and insurance organizations. These different arrangements each encourage physicians to treat otherwise similar patients in different ways. Fee-for-service insurance encourages the provision of costly services. Supply-side cost-sharing mechanisms, such as capitation payment, provide incentives to limit the volume and intensity of service (Randall P. Ellis and Thomas G. McGuire, 1993). Managed-care restrictions on the use of particular services or providers, through utilization review and gate-keeping, directly limit access to costly services. This last group of restrictions has been the focus of recent concerns that managed care reduces patient access to care by limiting the choice of physician.

While policy attention has focused on these demand limits, access to a choice of physicians is likely to be a consequence both of explicit insurance rules and of the willingness of physicians to supply services to patients with particular insurance arrangements. Physician decisions to accept multiple insurance arrangements will depend on the structure of physician practice. If physicians can costlessly vary their practice patterns from one patient to the next, they will be willing to see patients with many different types of coverage. If, however, some components of practice that are differently affected by different reimbursement incentives are fixed across patients, multi-payer practices will be more costly to operate than single-payer practices. Growing heterogeneity in payment mechanisms will encourage increasing specialization in physician practice arrangements. Without directly limiting patient or provider choice, heterogeneity in pay-

ment mechanisms may alter the treatment choices available to patients.

I. Fixed Costs in Physician Practice

Several aspects of physician practice that may be differently affected by different reimbursement incentives appear to be fixed across patients. For example, in 1995, the interquartile range of medical equipment expenditures within general practice was \$10,000 (American Medical Association, 1997). Incentives to use this equipment, and make these investments, vary with the level and nature of payment and with managed-care rules. Similarly, treatment patterns may be fixed, in part, at the practice level. For example, physicians may develop referral networks of specialists, they may select particular drugs to treat particular conditions, and they may organize visits in specific ways. Analyses of data on physician office practice suggest that the characteristics of other payers in a practice are significant determinants of patient treatment (Glied, 1998). Heterogeneity of payment types within a practice means that there is a less-than-optimal fit between practice characteristics and incentives for some patients.

II. Data

Data used in the analyses below are drawn from the 1991–1995 National Ambulatory Medical Care Survey (NAMCS), an annual nationally representative survey of randomly sampled physicians in office-based practice. Over 1,100 physicians were sampled in each survey year. The NAMCS asks physicians to report characteristics of about 30 randomly sampled visits that take place within a sample week. In total, the surveys include information on about 35,000 patient visits each year, including the expected source of payment for the visit, the geographic location of the visit, and the physician's specialty. (Information about coding procedures is available from the author

* School of Public Health and Department of Economics, Columbia University, 600 West 168th St., New York, NY 10032.

upon request.) The analyses below include only data on visits paid by Medicare (or Medicare HMO's) and by private payers (including HMO's). The NAMCS data do not contain detailed information about the type of payment. For example, the data do not indicate whether a physician received capitation or fee-for-service payment.

III. Heterogeneity and the Organization of Practice

Simple tabulations of the NAMCS data suggest that specialization by payer does, indeed, occur. In 1995, for example, 44 percent of Medicare patients seen by general internists were seen by physicians half or more of whose visits were paid by Medicare. Similarly, half of all private HMO visits to general practitioners were to physicians half or more of whose visits were paid by HMO's.

A more rigorous way of measuring this specialization is the index of dissimilarity.¹ Dissimilarity indexes are usually used to measure housing segregation within cities (see e.g., Douglas S. Massey and Nancy A. Denton, 1993; David M. Cutler and Edward L. Glaeser, 1997). In this context, the index of dissimilarity measures the percentage of Medicare patients (or private patients) who would have to switch physicians for the allocation of patients within each practice to be proportionate to that payer's share of the population. The analyses below examine dissimilarity indexes between private-pay patients under 65 and Medicare patients. The index is computed for each specialty, within each geographic region, in rural and urban areas separately (a total of 88 cells). The index of dissimilarity varies from 0 (no segregation) to 1 (complete segregation). I use a method suggested by Henry Inman and Edwin Bradley (1991) to compute the dissimilarity index that would be expected to arise under a random allocation in this sample.

Table 1 reports actual and expected dissimilarity indexes for selected specialties and for

TABLE 1—ESTIMATED AND EXPECTED DISSIMILARITY INDEXES FOR MEDICARE AND PRIVATE-PAY PATIENTS

Practice type	Estimated dissimilarity index	Expected dissimilarity index under random allocation
Specialties		
General or family practice	0.33	0.16
Internal medicine	0.35	0.15
Dermatology	0.22	0.15
Urology	0.27	0.17
Medical specialties	0.34	0.16
Surgical specialties	0.50	0.23
Areas		
Metropolitan area	0.35	0.23
Nonmetropolitan area	0.28	0.19

Notes: The expected dissimilarity index is computed as suggested in Inman and Bradley (1991). The standard deviation of the expected index is equal to or less than 0.02 in each case.

Source: National Ambulatory Medical Care Survey.

rural and urban areas for the 1995 NAMCS data.² The results suggest that there is substantially more segregation of patients into practices than would be expected under a random allocation of patients (the standard deviation of the expected dissimilarity index is always below 0.02). Such segregation at a point in time is not surprising. Medicare patients are older than private-pay patients, and the physicians who treat older patients may specialize in the care of conditions that affect the elderly. Consistent with this type of medical specialization, categories that contain heterogeneous specialties (other medical specialties and other surgical specialties) exhibit greater dissimilarity than more homogeneous categories. Similarly, opportunities for specialization by patient characteristics are greater in specialties such as general practice and internal medicine,

¹ Prior research in this area has focused on the choice of physician among patients with Medicaid coverage and has employed the Gini coefficient as an index of patient segregation (J. B. Mitchell and J. Cromwell, 1980).

² Data are weighted by the number of observations (total visits) used to construct the index. Several specialties are excluded from the analysis, either because they largely exclude either Medicare or private-pay patients or because the Physician Payment Review Commission did not provide specialty-specific payment information for them.

which serve large markets within a geographic region, and in urban areas.

Specialization related to patient health characteristics should be uncorrelated with changes in payment incentives and insurance structures. To measure whether practice is responsive to change in these incentives, I examine the relationship between these indexes and measures of payment heterogeneity over the period 1991–1995.³ I construct two measures of payment heterogeneity: HMO penetration and Medicare payment rate changes. NAMCS data are used to construct measures of private-sector and Medicare HMO penetration by region, rural or urban area, and specialty category, weighted by the NAMCS sample weights. The level of private-sector HMO penetration in the included specialties increased from 22 percent in 1991 to 27 percent in 1995, while the level of Medicare HMO penetration increased from 6 percent in 1991 to 15 percent in 1995. If HMO private-market penetration simply led physicians to substitute HMO dollars for fee-for-service dollars, while continuing to treat the same or similar patients, increases in the HMO share would lead to no change in the Medicare–private dissimilarity index. If Medicare patients were displaced in proportion to their share of the overall market, the coefficient on the private HMO share would be equal to 1.

The Physician Payment Review Commission has computed specialty-level changes in physician payments for 1991–1993 and 1994–1995 (Physician Payment Review Commission, various years). Data on evaluation and management and procedure-update levels, weighted by specialty-specific weights, were used to bridge the 1993 and 1994 data. The analyses include a separate trend for Medicare payments in 1994 and 1995 to adjust for error in this procedure.⁴ Over this period, fees

TABLE 2—THE EFFECT OF HMO SHARE AND MEDICARE FEES ON THE INDEX OF DISSIMILARITY BETWEEN MEDICARE AND PRIVATE PATIENTS

Independent variable	(i)	(ii)	(iii)	(iv)
HMO share of private visits	0.12* (0.05)	0.17* (0.06)		0.19* (0.06)
HMO share of Medicare visits		–0.17† (0.10)		–0.23* (0.08)
Medicare fee index			–0.004* (0.001)	–0.004* (0.001)

Notes: Results are from linear regressions, weighted by the number of observations used to construct the dissimilarity index. Standard errors are Huber-White adjusted. Regressions include region, rural, specialty, and year dummies. Analyses include 361 observations.

† Statistically significant at the 10-percent level.

* Statistically significant at the 5-percent level.

within and across specialties changed dramatically. Between 1991 and 1995, Medicare fees paid to general practitioners increased 35 percent, while those paid to cardiologists fell 5 percent. In 1991, Medicare fees averaged 65 percent of private rates (Physician Payment Review Commission, 1996). Thus, increases in Medicare fees would tend to reduce fee heterogeneity relative to the private market. If differences in fees encourage specialization, higher Medicare fees would lead to reductions in the dissimilarity index.

The results of analyses of the effect of measures of payment heterogeneity on indexes of dissimilarity are reported in Table 2. The analyses include specialty, region, rural, and year fixed effects, Huber-White robust standard errors are reported, and the analyses are weighted by the number of visits used to construct the dissimilarity indexes.⁵

Increases in the share of private-sector patient visits paid by HMO's lead to greater segregation of Medicare patients from private patients. As HMO penetration increases, HMO patients displace some potential fee-for-service Medicare patients. A 5-percentage-point

³ The ratio of Medicare patients to private patients in the sample remained roughly constant over this period, so that the indexes can be meaningfully compared across years of the survey.

⁴ The analyses were also conducted including an index of private fees based on physician visit fees for established patients (American Medical Association, 1997). Visit fees were only available for three specialties. This index was never significant, and inclusion of the index did not affect the results for the Medicare fee index.

⁵ The results reported in Table 2 are quite robust to alternative specifications. The results obtain for subperiods of the data, in both rural and urban areas, and are of a similar magnitude (although no longer statistically significant) in first differences. Similar results are also obtained by excluding all cells where fewer than 15 physician practices were used to construct the index and eliminating weighting.

increase in the share of private visits paid by HMO (as occurred between 1991 and 1995) raises the index by 1 percentage point (about 3 percent), significantly more than 0 but less than 1. An increase in the share of Medicare visits paid by HMO's reduces segregation of Medicare patients by about the same amount.

As expected, higher Medicare fees reduce dissimilarity. A 5-percent increase in Medicare fees leads to a 2-percentage-point decline in the dissimilarity index. Over the 1991–1995 period, patient segregation in general practices declined (substantially) while patient segregation in nine of the ten other specialties increased, leaving the overall level of patient segregation roughly constant.

IV. Implications

These results suggest that increasing payment heterogeneity leads to increased patient segregation. This finding has several implications. First, if practices become more concentrated, the effects of incentives of all types will grow stronger. For example, reductions in Medicare fee-for-service fees may generate larger volume responses for Medicare (McGuire and Mark V. Pauly, 1991). Similarly, the incentive to “manage care” in treating HMO patients is likely to grow if it is no longer offset by the incentives of treating fee-for-service patients.

Second, specialized suppliers (physicians) who have made purchaser-specific investments are vulnerable to opportunism by payers (Oliver E. Williamson, 1975). In other industries, purchaser-specific fixed costs of this type have led to long-term contracting and vertical integration (Williamson, 1975; Paul L. Joskow, 1993). As payment heterogeneity in the medical marketplace grows, vertical integration is likely to increase.

Finally, concerns about patient segregation, or access to a choice of physicians, are based on the assumption that patients with a choice among all physicians can choose those who are best. When insurance plans can be rank-ordered (e.g., by level of fee), segregation can be a measure of quality. For example, fee-for-service Medicaid patients have historically been segregated in practices served by physi-

cians with lower-than-average credentials (Mitchell and Cromwell, 1980). When, however, payment heterogeneity arises because of differences in the method of payment, the connection between quality and segregation is not as clear. Practices that specialize in the treatment of patients who pay capitation fees may be of better, worse, or equal quality to those that specialize in the treatment of fee-for-service patients. Payment heterogeneity, whatever its source, can generate increased segregation, but the quality implications of segregation differ according to the nature of the heterogeneity.

REFERENCES

- American Medical Association. *Physician marketplace statistics, 1996*. Chicago, IL: American Medical Association, 1997.
- Cutler, David M. and Glaeser, Edward L. “Are Ghettos Good or Bad?” *Quarterly Journal of Economics*, August 1997, 112(3), pp. 827–72.
- Ellis, Randall P. and McGuire, Thomas G. “Supply-Side and Demand-Side Cost Sharing in Health Care.” *Journal of Economic Perspectives*, Fall 1993, 7(4), pp. 135–51.
- Glied, Sherry. “Too Little Time? The Diagnosis and Treatment of Mental Health Problems in Primary Care.” *Health Services Research*, 1998 (forthcoming).
- Inman, Henry F. and Bradley, Edwin L., Jr. “Approximations to the Mean and Variance of the Index of Dissimilarity in $2 \times C$ Tables under a Random Allocation Model.” *Sociological Methods and Research*, November 1991, 20(2), pp. 242–55.
- Joskow, Paul L. “Asset Specificity and the Structure of Vertical Relationships: Empirical Evidence.” In Oliver E. Williamson and Sidney G. Winter, eds., *The nature of the firm: Origins, evolution, and development*. Oxford: Oxford University Press, 1993, pp. 117–37.
- Massey, Douglas S. and Denton, Nancy A. *American apartheid: Segregation and the making of the underclass*. Cambridge, MA: Harvard University Press, 1993.
- McGuire, Thomas G. and Pauly, Mark V. “Physician Responses to Fee Changes with

Multiple Payers." *Journal of Health Economics*, 1991, 10(4), pp. 385-410.

Mitchell, J. B. and Cromwell, J. "Medicaid Mills: Fact or Fiction." *Health Care Financing Review*, Summer 1980, 2(1), pp. 37-49.

Physician Payment Review Commission. *Annual report*. Washington, DC: U.S. Government Printing Office, various years.

Williamson, Oliver E. *Markets and hierarchies: Analysis and antitrust implications*. New York: Free Press, 1975.

What Has Increased Medical-Care Spending Bought?

By DAVID M. CUTLER, MARK McCLELLAN, AND JOSEPH P. NEWHOUSE*

As is well known, medical-care spending has increased dramatically for many decades. Although some popular discussions date the beginning of the increase to the mid-1960's with the passage of Medicare and Medicaid, in fact real per-person health-care spending has risen roughly 4 percent per year in each decade since the 1940's (Table 1). As a result of the more than half-century of growth, real per-person medical-care spending is now more than 11 times what it was in 1940. Moreover, the spending increase is not specific to the United States. Annual rates of increase for the G-7 countries from 1960 to 1990, for example, are shown in Table 2.

Several authors have advanced the notion that this sustained rate of increase can be proximately accounted for by the increased capabilities of medicine (William Schwartz, 1987; Henry Aaron, 1991; Burton Weisbrod, 1991; Newhouse, 1992). Because there is no straightforward summary measure of these capabilities, this hypothesis cannot be conclusively tested. Nonetheless, it has now achieved a degree of acceptance, in part because it is difficult to think of another factor that is common across six consecutive decades and across many countries with different health-care financing institutions. Here we accept for the sake of argument that a substantial portion of the spending rise is attributable to the increased capabilities of medicine and ask what the spending increases—and inferentially the increased capabilities—have bought.

An economist with no exposure to health care might ask why one should care about this question. After all, one does not hear the

analogous question of what increased spending on fax machines has bought. Some would answer the question of why one should care by rejecting the applicability of consumer sovereignty in medical care, either because insurance distorts private demands, physicians and other medical-care providers are imperfect agents, or externalities, or all of the above (e.g., Robert Evans, 1984). Here we point to a different reason: an answer is necessary to derive a proper cost-of-living index.

I. A Common View of the Answer

Whatever the reason for wanting an answer, many think they know it, namely, that the increased spending has bought little because the marginal unit of spending is heavily subsidized by insurance. And a good bit of evidence seems to back the claim that the marginal unit of medical care has little measurable effect on health, or that nonmedical factors have a much larger effect, or both. One example is Victor Fuchs's (1974) justly famous "tale of two states," comparing 1960's mortality rates in Nevada and Utah. Infant mortality rates were 40-percent higher in Nevada, and mortality rates at most ages for both males and females were similarly higher (e.g., for 40–49-year-old males and females they were 54- and 69-percent higher in Nevada). Because the states were similar in income, environment, and the availability of medical care, Fuchs attributed these mortality differences to life-style differences between the two states. This telling anecdote established in the minds of many health economists that "it's lifestyle, stupid" and inferentially that "it's not medical care." Subsequently the Rand Health Insurance Experiment showed little or no effect on the health outcomes of the average person from the additional medical care induced by free care (Newhouse and the Insurance Experiment Group, 1993).

Others note that the century-long downward trend in mortality showed little change after

* Cutler: Department of Economics, Harvard University, Cambridge, MA 02138, and National Bureau of Economic Research; McClellan: Department of Economics, Stanford University, Stanford, CA 94304-6072, and National Bureau of Economic Research; Newhouse: Division of Health Policy Research and Education, Harvard University, 180 Longwood Avenue, Boston, MA 02115.

TABLE 1—REAL PER-PERSON GROWTH IN PERSONAL HEALTH-CARE SPENDING, UNITED STATES, 1930–1995

Decade	Annual real growth per capita (percent)	Cumulative real per-person spending at end of decade (1940 = 100)
1930's ^a	1.4	100
1940's	4.0	148
1950's	3.6	211
1960's	6.1	381
1970's	4.5	592
1980's	5.2	983
1990–1995	3.4	1161

Sources: Newhouse (1992) for 1930–1960; *Health Care Financing Review* (Fall 1996) for 1960–1995. Population is taken from the *Statistical Abstract of the United States*. Spending is deflated by the GDP deflator, taken from the *Economic Report of the President*.

^a 1929–1940.

the advent of antibiotics (Thomas McKeown, 1976; Robert Fogel, 1994; Samuel Preston, 1996). These authors emphasize the importance for mortality of nutrition and of public-health measures, especially the safety of the water supply, rather than personal medical-care services.

However, neither the theory based on the subsidy from insurance nor the evidence just cited justifies the conclusion that the increased spending over time had little benefit. Even within a static context, inferences about moral hazard are difficult to make. Our intuition about the usual Harberger triangle (Mark Pauly, 1968) comes from situations in which compensated and uncompensated demand curves do not much differ. In health care, however, many interventions for low-probability events may be expensive relative to income. As a result, even if insurance provided lump-sum benefits, consumers could well want income transferred to sick states (David de Meza, 1983; Pauly, 1983). Thus, much of the increased spending on medical care induced by insurance may represent efficient spending, precisely because consumers may value income extremely highly when sick. Therefore, spending of insured consumers will exceed (potentially by a great deal) the spending of noninsured consumers, but the increase need not represent a welfare loss.

TABLE 2—ANNUAL RATE OF REAL PER-PERSON INCREASE IN MEDICAL-CARE SPENDING, G-7, 1960–1990

Country	Annual rate of increase (percent)
Canada	4.7
France	5.5
Germany	4.4
Italy	6.1
Japan	8.2
United Kingdom	3.7
United States	4.8

Notes: The 4.8 value given for the United States differs from the 5.2-percent growth rate implied by Table 1 because of revisions that the Health Care Financing Administration subsequently made to the U.S. rates of spending and because of different definitions of health-care spending. Spending is deflated by the GDP deflator.

Source: *Health Care Financing Review* (Summer 1992).

In a dynamic context, the evidence that the marginal value of medical care at a point in time is low does not imply that the average value of medical-technology changes over several decades is low. To measure cost-of-living indexes accurately, however, one needs to know the average value of medical-technology changes.

Accurate indexes are important for at least three reasons. First, as is well known, medical care is almost 14 percent of the U.S. GDP and is typically between 6 percent and 10 percent of GDP in most developed countries. At this size, substantial errors in the medical-care index can have nontrivial effects on measured economy-wide productivity. Second, there are numerous proposals to convert the Medicare program to a defined contribution program (e.g., Henry Aaron and Robert Reischauer, 1995). If, however, an increasing burden were not to be shifted to current beneficiaries, such a proposal would require a reliable medical-care price index, analogous to the consumer price index (CPI) for adjusting Social Security or wages. Such a price index would not necessarily insulate beneficiaries from cost increases associated with beneficial new services; how the costs of such change would be distributed would require a further policy decision. Third, and more subtly, the upward biases that many believe have

historically existed in the medical-care component of the CPI give an impression that spending increases largely stem from increases in unit prices, which in turn may be perceived as increases in rents. This has generated periodic calls for medical-care price controls, as for example the treatment of pharmaceuticals in the Clinton Health Security Act.

II. Deriving a Cost-of-Living Index for Medical Care

The most difficult problem in deriving a medical care cost-of-living index is the treatment of new or improved goods—the pricing of changes in medical technology. Existing medical-care price indexes typically link a new medical-care good or service into the index in a way that does not alter the value of the index at the time of introduction. This was never a very satisfactory procedure because it ignored the potential increase in welfare from a costly but beneficial new product (e.g., non-invasive imaging). Arguably it will be even less satisfactory in the future if observed prices relate to premiums of managed-care plans. As new but costly technologies are introduced, health-plan premiums will presumably increase, and this could be inappropriately treated as a pure price increase rather than a quality improvement. The reverse situation might occur if managed-care companies reduced moral hazard by reducing services and thus lowered premiums. Not all of the premium decrease would necessarily be a pure price decline.

Two approaches to new goods have been taken in the price-index literature. Hedonic analysis is a standard approach (Zvi Griliches, 1971), but the widespread nature of administered prices and the problem of identification make it difficult to apply in medical care. As an example, consider the treatment of heart attacks. An important and expensive component of heart-attack treatment is care in an intensive-care unit, but many types of patients with differing needs for nursing time are treated in intensive-care units. Because of uniform pricing per day across patients, there will be cross subsidies among patients with different diagnoses, and the size of the cross subsi-

dies could well change over time.¹ In addition, hedonic analysis relies on consumers making price and quality decisions, but this is typically not the case in medical care. Secondly, one can also specify a model of medical decision-making (Franklin Fisher and Griliches, 1995), but there is no agreement on how to do this. An alternative to inferring the value of new capabilities from observed prices is to measure the output of medical care in physical terms, such as mortality reduction, and then value that change directly. We and Dahlia Remler have done this for the treatment of acute myocardial infarction (AMI) or heart attack among the elderly in the United States between 1984 and 1991 (David Cutler et al., 1996).² Heart attacks provide a good example for at least two reasons: (i) they have a defined and observable beginning; and (ii) reductions in mortality are an important outcome of heart-attack treatment, so valuation techniques can borrow from the substantial value-of-life literature (W. Kip Viscusi, 1993).

We found substantial upward bias when comparing a cost-of-living index to standard price indexes for heart attacks. Part of the bias came from the infrequency of updating the market basket of goods in the traditional CPI and the rapidity of change in inputs used to treat heart attacks. But part of the bias occurred because the measured index ignored the reduction in mortality rates that occurred over the period we studied. As a result of this reduction, life expectancy among elderly heart-attack victims increased 13 percent between 1984 and 1991, from 5 years and 2 months to 5 years and 10 months.

Of course, all the gain cannot necessarily be attributed to improved medical treatment, even using a broad definition of medical treatment such as faster response times and better-equipped ambulances. Improvements in the treatment of other diseases that heart-attack

¹ We presume that because of administered pricing within the hospital the price of a day in the intensive-care unit does not necessarily reflect marginal cost and that the markup may change over time.

² We limited our study to the elderly because of the availability of near-universal data from the Medicare program.

patients might also have (e.g., diabetes) would affect life expectancy, but the change would not be attributable to the monies spent on heart attacks. Improvements in lifestyle, such as better diet or increased exercise, would also prolong life following a heart attack and may change the initial severity of the attack (in either direction). We could not hope to control for these additional factors. As a rough correction, we calculated the gain in life expectancy among all the elderly, about four months. Using a difference-in-difference procedure, we attributed half the gain in life expectancy, or four of the eight months, to the improved treatment of heart attacks. The attribution of about half the gain to treatment is remarkably close to an estimate of the proportion of the decrease in mortality from all coronary heart disease (not just AMI) that should be attributed to improved intensive treatment, derived using a much different and more detailed method (Maria Hunink et al., 1997). Valuing the four-month gain in life expectancy at a relatively conservative \$25,000 per life-year implies that the measured cost-of-living index for heart-attack treatment should be about 4 percentage points less than if one ignores the change in life expectancy. Because this figure does not consider changes in the quality of life, it probably understates the benefits of the change in treatment.³

III. Conclusions

Accurate cost-of-living indexes require valuing the enhanced capabilities of medicine, just as the indexes must account for the enhanced capabilities of personal computers. Unlike personal computers, however, it is hard to apply hedonic analysis to medical care. It is thus important to carry out studies similar to our heart-attack study for diseases other than heart attacks. Heart attacks are a particularly favorable case; other advances will have their primary effect on functional capacity or quality of life (e.g., improved intraocular lenses for cataracts; laparoscopic surgical techniques that involve less pain

and more rapid recovery times; antidepressant drugs with fewer side effects). Our ability to measure and value such outputs is improving (F. Reed Johnson et al., 1997) but is clearly not yet at the same level as our ability to value changes in mortality.

REFERENCES

- Aaron, Henry. *Serious and unstable condition: Financing America's medical care system*. Washington, DC: Brookings Institution, 1991.
- Aaron, Henry and Reischauer, Robert. "The Medicare Reform Debate: What Is the Next Step?" *Health Affairs*, Winter 1995, 14(4), pp. 8-30.
- Cutler, David M.; McClellan, Mark; Newhouse, Joseph P. and Remler, Dahlia K. "Are Medical Prices Declining?" National Bureau of Economic Research (Cambridge, MA) Working Paper No. 5750, September 1996.
- de Meza, David. "Health Insurance and the Demand for Medical Care." *Journal of Health Economics*, March 1983, 2(1), pp. 47-54.
- Evans, Robert G. *Strained mercy*. Toronto: Buttersworth, 1984.
- Fisher, Franklin M. and Griliches, Zvi. "Aggregate Price Indices, New Goods, and Generics." *Quarterly Journal of Economics*, February 1995, 110(1), pp. 229-44.
- Fogel, Robert. "Economic Growth, Population Theory, and Physiology: The Bearing of Long-Term Processes on the Making of Economic Policy." *American Economic Review*, June 1994, 84(3), pp. 369-95.
- Fuchs, Victor R. "Some Economic Aspects of Mortality in Developing Countries," in Mark Perlman, ed., *The economics of health and medical care*. London: Macmillan, 1974; reprinted in Victor R. Fuchs, *The health economy*. Cambridge, MA: Harvard University Press, 1986, pp. 181-99.
- Griliches, Zvi, ed. *Price indices and quality change*. Cambridge, MA: Harvard University Press, 1971.
- Hunink, Maria G. M.; Goldman, Lee; Tosteson, Anna N. A.; Mittelman, Murray A.; Goldman, Paula A.; Williams, Lawrence W.; Tsevak, Joel and Weinstein, Milton C. "The Recent Decline in Mortality from Coronary Heart Disease." *Journal of the American Medical*

³ Specifically, there is some evidence of gain in functioning as well as reduction in mortality.

- Association*, 19 February 1997, 277(7), pp. 535-42.
- Johnson, F. Reed; Fries, Erin E. and Banzhaf, H. Spencer. "Valuing Morbidity: An Integration of the Willingness-to-Pay and Health-Status Index Literatures." *Journal of Health Economics*, 1997, 16(6), pp. 641-55.
- McKeown, Thomas. *The modern rise of population*. New York: Academic Press, 1976.
- Newhouse, Joseph P. "Medical Care Costs: How Much Welfare Loss?" *Journal of Economic Perspectives*, Summer 1992, 6(3), pp. 3-21.
- Newhouse, Joseph P. and the Insurance Experiment Group. *Free for All? Lessons from the RAND health insurance experiment*. Cambridge, MA: Harvard University Press, 1993.
- Pauly, Mark V. "The Economics of Moral Hazard." *American Economic Review*, June 1968, 58(3), pp. 231-37.
- . "More on Moral Hazard." *Journal of Health Economics*, March 1983, 2(1), pp. 81-86.
- Preston, Samuel H. "American Longevity: Past, Present, and Future." Syracuse University Maxwell School Policy Brief No. 7, 1996.
- Schwartz, William B. "The Inevitable Failure of Cost Containment Strategies: Why They Can Provide Only Temporary Relief." *Journal of the American Medical Association*, 9 January 1987, 257(2), pp. 220-24.
- Viscusi, W. Kip. "The Value of Risks to Life and Health." *Journal of Economic Literature*, December 1993, 31(4), pp. 1912-46.
- Weisbrod, Burton. "The Health Care Quadrilemma: An Essay on Technological Change, Insurance, Quality of Care, and Cost Containment." *Journal of Economic Literature*, September 1991, 29(3), pp. 523-52.

SOCIAL SECURITY AND THE REAL ECONOMY: EVIDENCE AND POLICY IMPLICATIONS[†]

Social Security: Privatization and Progressivity

By LAURENCE J. KOTLIKOFF, KENT A. SMETTERS, AND JAN WALLISER *

The Balanced Budget Agreement of 1997 achieved budget balance by the year 2002 but did not resolve the nation's long-term fiscal problems. Those problems stem, in large part, from the Medicare and Social Security programs whose projected expenditures outstrip their projected receipts. One government commission is now dealing with Medicare's finances. Another, the Social Security Advisory Council, recently delivered three mutually exclusive sets of recommendations, one of which calls for the system's privatization.

A sizable literature shows that privatizing Social Security would increase the economy's long-run productive capacity at the price of higher fiscal burdens for those generations alive during the transition.¹ Less well explored is how Social Security's privatization would alter the intragenerational resource distribution. The answer, as discussed below, depends on the method of privatization. Some privatization schemes entail no progressive elements. But even these schemes may help the long-run poor more than the long-run rich. Other schemes use flat minimum benefits or progressive contribution matches to enhance intragenerational progressivity. This paper studies the interface of Social Security privat-

ization and progressivity using a large-scale overlapping-generations model. The paper's bottom line is that privatization and progressivity can be mutually compatible, particularly if the redistribution is achieved through progressive matching of individual accounts and is financed with a consumption tax.

I. The Model

Our model (Kotlikoff et al., 1997) is a substantially enhanced version of the dynamic general-equilibrium model of Alan Auerbach and Kotlikoff (1987). It features 55 overlapping generations and solves for the economy's perfect-foresight transition path. The model, which is calibrated to U.S. data, has three sectors: households, firms, and the government. Households allocate their full lifetime resources to consumption and leisure over their life span (ages 21–75); retirement decisions are endogenous. Household behavior is governed by a time-separable constant-elasticity-of-substitution utility function with intertemporal and intratemporal elasticities of substitution of 0.25 and 0.8, respectively. Population and productivity both grow at 1 percent per year.

The model follows Don Fullerton and Diane Lim Rogers (1993) in capturing intragenerational heterogeneity by dividing each cohort into 12 lifetime-earnings classes. The 12 classes represent the 10 deciles of the population ranked by lifetime income where the bottom and top deciles are each divided into percentiles of 2 and 8 percent. Age-productivity profiles for each of these classes were estimated from the Panel Study of Income Dynamics. Firms are perfectly competitive and maximize profits subject to a Cobb-Douglas production function with a capital-income share of 25 percent.

[†] *Discussants:* Barry Bosworth, Brookings Institution; Martin Feldstein, Harvard University; Theodore Bergstrom, University of Michigan.

* Kotlikoff: Department of Economics, Boston University, 270 Bay State Road, Boston, MA 02215, and NBER; Smetters and Walliser: Congressional Budget Office, U.S. Congress, Washington, DC 20515. The views herein do not necessarily reflect those of the CBO. We thank Joyce Manchester for helpful comments.

¹ The literature includes Martin Feldstein and Andrew Samwick (1998), Patricio Arrau and Klaus Schmidt-Hebbel (1993), Bernd Raffelhüschen (1993), Kotlikoff (1996), and Kotlikoff et al. (1997).

The government collects taxes to finance government consumption and OASDI benefits. In the initial steady state, the U.S. tax system is represented by a proportional state and federal consumption tax, a proportional state income tax, a proportional federal capital income tax, a proportional federal payroll tax with an earnings ceiling, and a progressive federal wage tax with a standard deduction.² Thus, households' budget constraints are non-convex and nondifferentiable because of the payroll-tax ceiling and the wage deduction. The model's initial OASI payroll tax rate is 9.9 percent. The total payroll tax rate including hospital insurance (HI) and disability insurance (DI) is 14.7 percent.³

The model applies Social Security's OASI inflation-indexed benefit formula to each agent's average indexed earnings. Because about 50 percent of Social Security benefits are paid to survivors and spouses, we multiply benefits by a factor of 2. All workers are covered by Social Security, and all correctly perceive and value the marginal OASI benefits received when they earn an extra dollar. Our model incorporates many complex aspects of the economy but ignores Social Security's role in inter- and intragenerational risk-pooling.⁴

II. Privatizing Social Security

Privatizing Social Security involves three elements: (i) forcing workers to contribute to personal accounts, (ii) honoring accrued benefits, and (iii) choosing a method to finance accrued benefits in the transition.⁵ Since agents in our model are free to borrow against mandatory accounts, there is no need to add a private pension system. Instead, privatization is

achieved by simply eliminating the OASI payroll tax and phasing out OASI benefits starting 10 years after the reform begins. The decade wait to reduce benefits allows current retirees to get full benefits. After the 10th year, benefits are reduced by 2.2 percent per year for the next 45 years. This gradual phase-out of benefits captures the provision of accrued benefits to existing workers.

Three alternative tax regimes—a payroll tax, an income tax, and a consumption tax—are used to finance accrued OASI benefits. In each simulation, federal wage and capital income taxes are adjusted endogenously to balance the federal budget. Specifically, these taxes are adjusted so that the average tax rate on wage income changes by the same percentage as those on capital income. Other adjustments of the tax schedule that would make privatization more progressive could be considered.

The top panel of Table 1 shows how privatization alters the economy. All three simulations produce the same long-run steady state. Compared to the initial steady state, the long-run steady state features a 39-percent larger capital stock, a 5-percent larger supply of labor, and a 13-percent larger level of output.

However, important differences exist in the speed of the transition. Financing the transition with a payroll tax reduces aggregate labor supply initially because privatization removes the link between taxes and benefits for young workers. Thus, output is slightly lower in the first decade. Using income-tax finance reduces both labor supply and capital accumulation in the short run, further reducing the speed of transition. A transition financed by a consumption tax, on the other hand, can encourage savings and labor supply by taxing existing non-Social Security wealth, leading to growth even in the short to medium term.

Privatization increases the welfare of future workers, with the largest gains accruing to the average worker (see the first panel of Table 2). Utility, measured in wealth equivalents, would rise by 8 percent for average earners, 6 percent for the poorest agents, and 4.4 percent for the richest agents. While the poor benefit from reduced payroll taxes, their welfare is not affected by the growth-induced fall in income-

² The proportional capital income tax and the progressive wage tax approximate the federal income tax.

³ HI benefits are modeled as a constant transfer to agents of age 65 and over, and DI benefits are modeled as a constant transfer to agents below age 65.

⁴ Henning Bohn (1997) considers aggregate risk. He Huang et al. (1997) consider idiosyncratic earnings and longevity uncertainty. Walliser (1997) discusses the impact of privatization on annuities markets.

⁵ We do not address the actuarial value of performance guarantees. Those guarantees may create large unfunded liabilities in the case of rate-of-return uncertainty (Smetters, 1997).

tax rates because their earnings are exempt from taxation. Higher earners benefit less than average earners because most of their earnings are above the payroll-tax ceiling.

Welfare effects for generations alive during the transition depend heavily on the method of transition finance. Payroll-tax finance would harm those with the largest payroll-tax burden as a percentage of income: the lifetime poorest. The elderly are largely unaffected while workers carry most of the load. Income-tax finance would put the largest burden on high earners whose capped payroll taxes are replaced by a progressive wage tax and a proportional capital income tax. Consumption taxes would levy a tax on owners of non-Social Security wealth, placing a higher burden on older and middle-aged generations than payroll-tax finance. Consumption taxes place a burden on the poor that is similar to payroll taxes.

III. Privatizing with a Flat Minimum Benefit

Some plans, most prominently the Personal Security Account Plan of the Social Security Advisory Council, propose a pay-as-you-go-financed flat minimum benefit. We investigate this policy by (i) providing a wage-indexed flat minimum annual benefit of \$6,000 in the long run and (ii) paying a weighted average of the old OASI and the new flat minimum benefit during the transition. We consider the same three financing methods but assume that these alternative taxes also finance the flat minimum benefit. In this case, the different financing methods imply different steady states.

Providing a flat minimum benefit substantially reduces the output effect of privatization (see the second panel in Table 1). Long-run increases in capital, labor, and output are between 40 percent (consumption-tax finance) and 70 percent (income-tax finance), smaller than under complete privatization. Short-run effects on capital, labor, and output are similar to complete privatization but are substantially diminished by year 25.

There are two reasons for these outcomes. First, the continuing unfunded liability, which amounts to about half of the current unfunded liability in the Social Security system, reduces the effect of privatization on saving and capital

TABLE 1—MACROECONOMIC RESPONSES TO PRIVATIZATION (PERCENTAGE CHANGE FROM STEADY STATE)

		Year of transition			
Finance	Variable	5	10	25	150
Privatization:					
Payroll tax	K	0.3	0.7	5.2	38.6
	L	-1.1	-1.1	1.8	5.2
	Y	-0.8	-0.7	2.6	12.7
Income tax	K	-2.4	-5.0	-4.6	38.6
	L	-4.5	-4.7	0.0	5.2
	Y	-4.0	-4.8	-1.0	12.7
Consumption tax	K	1.8	4.1	12.8	38.6
	L	0.3	0.4	2.4	5.2
	Y	0.6	1.3	4.9	12.7
Privatization with Flat Minimum Benefit:					
Payroll tax	K	0.0	0.0	2.0	19.0
	L	-1.3	-1.4	0.2	2.3
	Y	-1.0	-1.1	0.6	6.2
Income tax	K	-2.8	-5.7	-8.7	12.4
	L	-4.7	-4.9	-2.9	1.2
	Y	-4.2	-5.1	-4.4	3.9
Consumption tax	K	1.4	3.2	8.9	23.2
	L	0.0	0.1	1.1	2.7
	Y	0.4	0.8	3.0	7.5
Privatization with Progressive Matching:					
Payroll tax	K	-0.7	-1.4	0.9	35.4
	L	-3.2	-3.3	-0.2	4.0
	Y	-2.6	-2.9	0.1	11.1
Income tax	K	-3.4	-7.1	-9.7	35.4
	L	-6.7	-7.3	-3.0	4.0
	Y	-5.9	-7.2	-4.7	11.1
Consumption tax	K	1.8	4.1	13.0	39.8
	L	-0.5	-0.4	1.7	4.5
	Y	0.0	0.7	4.4	12.4

Notes: K = capital stock, L = labor supply, Y = output; all runs assume positive marginal link between taxes and benefits.

accumulation. Second, the tax that finances the flat minimum benefit is now completely distortionary since benefits no longer change at the margin.

The long-run welfare gains, though much smaller, are more progressive than those under complete privatization, especially with income-tax financing. Also, because implicit debt is reduced only by half, transitional welfare losses tend to be smaller.

TABLE 2—CHANGES IN REMAINING LIFETIME UTILITY
(WEALTH EQUIVALENTS) AFTER PRIVATIZATION
(PERCENTAGE CHANGE FROM STEADY STATE)

Finance	Class	Year of birth			
		-54	-25	1	150
Privatization:					
Payroll tax	1	0.0	-2.0	-0.6	6.0
	6	-0.1	-1.4	-0.2	8.0
	12	-0.1	-0.6	-0.1	4.4
Income tax	1	-0.1	-0.2	3.2	6.0
	6	-1.3	-2.1	0.7	8.0
	12	-1.7	-3.6	-3.0	4.4
Consumption tax	1	0.7	-2.1	0.5	6.0
	6	-0.9	-1.7	1.6	8.0
	12	-1.5	-2.5	-1.0	4.4
Privatization with Flat Minimum Benefit:					
Payroll tax	1	0.0	-0.1	0.3	4.0
	6	-0.1	-0.8	-0.2	4.3
	12	-0.1	-0.5	-0.2	2.3
Income tax	1	-0.1	1.8	4.3	5.7
	6	-1.3	-1.6	0.6	4.4
	12	-1.7	-3.6	-3.4	0.5
Consumption tax	1	0.7	-0.3	1.4	4.9
	6	-0.9	-1.3	1.6	5.4
	12	-1.5	-2.5	-1.2	2.0
Privatization with Progressive Matching:					
Payroll tax	1	0.0	-1.0	1.6	8.0
	6	-0.4	-1.6	0.0	8.1
	12	-0.5	-1.7	-1.2	3.5
Income tax	1	-0.2	0.9	5.3	8.0
	6	-1.6	-2.4	0.7	8.1
	12	-2.1	-4.9	-4.3	3.5
Consumption tax	1	0.9	-1.6	1.9	7.6
	6	-1.1	-1.9	1.9	8.4
	12	-1.8	-3.2	-1.6	4.0

Notes: 1 = bottom 2 percent, 6 = fifth decile, 12 = top 2 percent of lifetime income distribution.

IV. Privatizing with Progressive Matching

Our final set of experiments considers matching contributions to mandatory private saving accounts in a progressive way. The government's match is calculated as a function of labor income, and it falls steadily as a percentage of earnings, starting at about 5 percent for the poorest. In absolute terms, it increases from about \$470 at annual earnings of \$10,000 to around \$840 for annual earnings of \$21,000 and stays constant thereafter. On a lifetime ba-

sis, the match provides a transfer to the poorest that exceeds the flat minimum benefit of the previous section by 30 percent. Workers fully incorporate the marginal subsidy associated with the progressive-contribution match into their decisions.

The first two runs reported in the third panels of Tables 1 and 2 finance the revenue shortfall from the tax credit by raising income taxes, and the third run raises consumption taxes. OASI benefits are phased out as above and financed with either a payroll tax, an income tax, or a consumption tax.

Progressive matching has a much less detrimental effect on growth than does a flat minimum benefit. The negative effects of increased income-tax rates (runs with payroll- and income-tax finance) or consumption-tax rates are substantially smaller than with a flat minimum benefit. However, since the government must finance all accrued benefits and the matching contribution during the early years, the transition with payroll-tax finance and income-tax finance is slower than in runs without matching. That slowdown is largely caused by the negative impact on aggregate labor supply due to initially higher progressive income taxes (runs with payroll- and income-tax finance) and the income effect from the matching of labor income. With consumption-tax finance of the matched contribution as well as accrued benefits, the transition path to the final steady state as well as to the final steady state itself is quite similar to that without the match.

The progressive matching of labor income leads to about the same percentage increase in welfare for low earners and average earners in the long run. Both groups fare as well as mean earners under complete privatization without matching. The lifetime richest, however, lose due to the progressivity of the match. Generations alive during the transition face a higher burden than under the flat minimum-benefit run; however, the lifetime poorest are better off than under privatization without matching.

V. Conclusions

Privatization can offer substantial long-run economic gains. But those gains are not free, nor are they immediate. Some transition gen-

erations face higher fiscal burdens, and depending on how the transition is financed, it can be quite slow.

Enhancing progressivity in a privatized system with a pay-as-you-go-financed flat minimum benefit comes at the cost of substantially smaller long-run macroeconomic and welfare gains. Matching workers' contributions on a progressive basis is an alternative means of making Social Security's privatization more progressive. Relative to a flat minimum benefit, this policy achieves an equally progressive long-run distribution of welfare. But it affords much larger long-run levels of capital, labor supply, output, and welfare.

REFERENCES

- Arrau, Patricio and Schmidt-Hebbel, Klaus. "Macroeconomic and Intergenerational Welfare Effects of a Transition from Pay-as-You-Go to Fully Funded Pensions." Working paper, Policy Research Department, World Bank, Washington, DC, 1993.
- Auerbach, Alan and Kotlikoff, Laurence J. *Dynamic fiscal policy*. Cambridge: Cambridge University Press, 1987.
- Bohn, Henning. "Social Security Reform and Financial Markets." Mimeo, University of California-Santa Barbara, 1997.
- Feldstein, Martin S. and Samwick, Andrew A. "The Transition Path in Privatizing Social Security," in Martin S. Feldstein, ed., *Privatizing Social Security*. Chicago: University of Chicago Press, 1998 (forthcoming).
- Fullerton, Don and Rogers, Diane Lim. *Who bears the lifetime tax burden?* Washington, DC: Brookings Institution, 1993.
- Huang, He; Imrohoroglu, Selahattin and Sargent, Thomas. "Two Computational Experiments to Fund Social Security." *Macroeconomic Dynamics*, February 1997, 1(1), pp. 7-44.
- Kotlikoff, Laurence J. "Privatizing Social Security: How It Works and Why It Matters," in James M. Poterba, ed., *Tax policy and the economy*, Vol. 10. Cambridge, MA: MIT Press, 1996, pp. 1-32.
- Kotlikoff, Laurence J.; Smetters, Kent A. and Walliser, Jan. "Opting Out of Social Security." Mimeo, Congressional Budget Office, Washington, DC, November 1997.
- Raffelhüschen, Bernd. "Funding Social Security Through Pareto-Optimal Conversion Policies." *Journal of Economics/Zeitschrift für Nationalökonomie*, Supplement 1993, 7, pp. 105-31.
- Smetters, Kent A. "Privatizing Social Security in the Presence of a Performance Guarantee." Mimeo, Congressional Budget Office, Washington, DC, October 1997.
- Walliser, Jan. "Understanding Adverse Selection in the Annuities Market and the Impact of Privatizing Social Security." Congressional Budget Office (Washington, DC) Technical Paper No. 1997-4, August 1997.

Perspectives on the Social Security Crisis and Proposed Solutions

By KEVIN M. MURPHY AND FINIS WELCH*

Many economists as well as a majority of the general public are pessimistic about the future of the U.S. Social Security System. Such pessimism is justified. Current estimates are that the system will run out of funds around 2030. This is bad news for today's young workers who will, even under current law, receive only about 50 cents in benefits for every dollar they will pay in Social Security taxes over their career (Murphy and Welch, 1996). The return these workers ultimately will receive will be even less, maybe much less, depending on what is done to balance the system, and when. With its low promised return and the uncertainty about its future, Social Security may seem more like social insecurity for many of today's workers.

The prospect of insolvency and overall dissatisfaction with the current system has led to the search for alternatives. Some call for privatization (Martin Feldstein and Andrew Samwick, 1996, 1997; Laurence Kotlikoff, 1998); some advocate a voluntary system while others favor less radical adjustments to the current system (Peter Diamond, 1996). In this paper we examine the current system, highlight the source of our current problems, and examine areas where the system can be improved through privatization or other adjustments. Our basic view is that many of the touted gains from privatization are more apparent than real, and any real gains have more to do with the details of what is done to the system (whether private or public) than with privatization per se. What appears to be most important is to do something about the problem sooner rather than later and to resolve the political uncertainty over what will be done to salvage the system.

* Graduate School of Business, University of Chicago, 1101 East 58th Street, Chicago, IL 60637, and Department of Economics, Texas A&M University, College Station, TX 77843-4228, respectively.

I. The Current System

The current Social Security System is essentially a tax and transfer program. Until 1986, tax rates were adjusted to keep current taxes and benefits roughly in balance so that intertemporal movements in resources were minimal (the system was essentially pay-as-you-go). Since 1986, the system has run a significant annual surplus which has ranged from \$23 billion to \$60 billion per year. Significant surpluses are expected to continue for the next 15–20 years until the peak of the baby-boom cohorts reaches retirement age. Accumulated system assets are invested in special U.S. Treasury securities. While the short-term financial picture is good, the long-term picture is not. As we mentioned in the Introduction, without any adjustments to tax or benefit formulas the system is expected to run out of funds around 2030.

Current law taxes wages at 10.7 percent up to the annual taxable maximum income (\$65,400 in 1997). Benefits are paid based on covered earnings (equal to earnings up to the tax maximum in each year) over an individual's career. Benefits are calculated by indexing earnings from each year of an individual's career to age 60 using the ratio of the economy-wide average earnings when the individual is age 60 relative to economy-wide average earnings in that year (earnings after age 60 are not indexed). Hence the rate of earnings growth for the economy determines the implicit preretirement return paid to workers for contributions under the system.

The 35 highest indexed earnings values are then used to compute average indexed monthly earnings (AIME) for the individual. The individual's primary insured amount, PIA (the benefit paid at normal retirement age) is a concave, piecewise linear function of the AIME amount with slopes 0.90, 0.32, and 0.15. The function for those becoming eligible

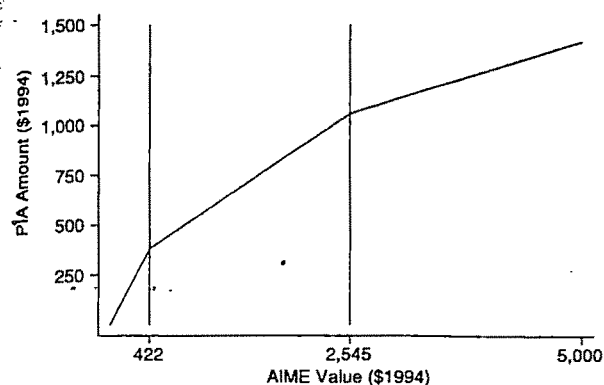


FIGURE 1. AIME AND SOCIAL SECURITY PIA AMOUNTS (AGE-65 BENEFIT)

for retirement (i.e., reaching age 62) in 1994 is illustrated in Figure 1.

An individual's benefits are then determined based on his or her PIA amount together with an actuarial adjustment for the age of retirement and an adjustment for CPI inflation since reaching age 62 according to

$$\text{Benefit}(a, k) = \theta_k \text{PIA} (\text{CPI}_a / \text{CPI}_{62})$$

where θ_k is the actuarial factor for a person retiring at age k . PIA is the individual's primary insured amount and $\text{CPI}_a / \text{CPI}_{62}$ measures the growth in prices between age 62 and the individual's current age, a . Currently, normal retirement age is 65, and θ_{65} is normalized to be 1.0. Under current law, θ_k is reduced by 1/180 for each month the individual retires prior to age 65 so that $\theta_{62} = 0.8$. Based on standard life tables and a 3.5-percent real rate of discount, this adjustment is close to the actuarially fair adjustments of 0.77 for men and 0.81 for women. Actuarial adjustments for those retiring after age 65 (1/200 per month) are much less than fair, with those retiring at age 69 receiving only 24-percent more than the PIA amount ($\theta_{69} = 1.24$), while the actuarially fair adjustments using a 3.5-percent real rate are 1.47 for men and 1.38 for women.

The returns embodied in current law are summarized in Table 1 which gives the present value of benefit payments per dollar of taxes paid (also in present value) for men and women who retire at ages 62 or 65 based on a typical lifetime work profile (see Murphy and Welch [1996] for details). The calculations

TABLE 1—MONEY'S-WORTH CALCULATIONS FOR PRIMARY BENEFITS BASED ON A 13.8-PERCENT OASI TAX RATE AND CURRENT LAW

Retirement category	Benefit segment			
	Second			Average, endpoint
	First	Marginal	Average, midpoint	
Men, retiring at age 65				
$r = 2.3$	1.04	0.37	0.56	0.48
$r = 3.5$	0.71	0.25	0.38	0.33
Men, retiring at age 62				
$r = 2.3$	1.06	0.38	0.57	0.49
$r = 3.5$	0.73	0.26	0.40	0.34
Women, retiring at age 65				
$r = 2.3$	1.30	0.47	0.71	0.60
$r = 3.5$	0.87	0.30	0.47	0.40
Women, retiring at age 62				
$r = 2.3$	1.30	0.47	0.70	0.60
$r = 3.5$	0.87	0.31	0.47	0.40
Third benefit segment				
Retirement category	Marginal	Average, midpoint	Average, tax maximum	
Men, retiring at age 65				
$r = 2.3$	0.17	0.38	0.33	
$r = 3.5$	0.12	0.26	0.22	
Men, retiring at age 62				
$r = 2.3$	0.18	0.39	0.33	
$r = 3.5$	0.12	0.26	0.23	
Women, retiring at age 65				
$r = 2.3$	0.22	0.47	0.41	
$r = 3.5$	0.14	0.31	0.27	
Women, retiring at age 62				
$r = 2.3$	0.22	0.47	0.41	
$r = 3.5$	0.15	0.32	0.27	

Notes: Marginal and average returns are equal throughout the first segment. Indexed earnings are year-of-retirement national-average wage-indexed values. As of 1994, the first segment ranged from zero to indexed average annual earnings of (approximately) \$5,000. The second segment extends to \$30,000, and the third extends to \$60,000. The midpoint of the second segment is \$17,500, and the midpoint of the third is \$45,000. Assumptions: (i) 0.5-percent real wage growth and 2.8-percent inflation with alternative real discount rates of 2.3, 3.0, and 3.5 percent; (ii) age-wage gradient corresponds to CPS average, 1967–1994; (iii) labor-force participation is part-time, 12 weeks, ages 16–19; part-time, 52 weeks, ages 20–24; and full-time, 52 weeks, age 25–retirement.

assume an OASI tax rate of 13.8 percent (a rate roughly equal to that required to finance the system in the long term) and use alternative real discount rates of 2.3 percent (the rate used by the Social Security actuaries) and 3.5 percent. The returns are calculated for various points along the benefit schedule and in both marginal and average terms. Most workers would fall between the midpoint of the second segment and the midpoint of the third segment

where average returns range from 26 to 57 cents on the dollar for men and from 31 to 71 cents on the dollar for women. Marginal returns are significantly lower, with men earning about 25 cents per dollar on the second segment and only 12 cents per dollar on the third. Women do only slightly better, with marginal returns of 30 and 15 cents per dollar on the second and third segments, respectively. Average returns for those at the tax maximum are only 22–33 cents on the dollar for men and 27–41 cents on the dollar for women. Clearly, returns are not very good.

The provisions outlined above (with the exception of the 13.8-percent tax rate) apply to all workers who have retired in recent years. This same basic system is scheduled to continue (with some adjustments to the normal retirement age and actuarial factors) for the indefinite future. While the benefit schedules will remain constant, the implied returns earned per dollar contributed will differ significantly across cohorts of workers since tax rates have been adjusted upward over time in a series of steps from 4.0 percent in 1955 to a high of 11.2 percent in 1990, with minor subsequent adjustments. In all likelihood, tax rates will be adjusted upward by a minimum of 2–3 percentage points (to 13.8 percent in our example) in the coming decades. The taxable maximum has also been increased over time, from \$4,200 in 1955 to \$65,400 in 1997 and is scheduled to continue to grow at the rate of overall earnings. Panel A of Table 2 illustrates the effects of these changes in tax rates and maximum taxable income figures by calculating the amount of tax dollars paid per dollar of AIME for workers aged 35, 45, 55, and 65 in 1997. The calculations use the historical tax rate series through 1997 and a 13.8-percent tax rate (the rate required to balance the system) from 1998 onward. Tax contributions per dollar of AIME are presented for workers earning different fractions of the average annual earnings over their careers. As the table illustrates, younger cohorts must contribute up to 67 percent more (25.10 vs. 15.25) per dollar of AIME than did those retiring at age 65 in 1997. However, this is only part of the story. As panel B illustrates, the increase in tax rates and taxable earnings through time implies that young cohorts must make larger absolute tax

TABLE 2—SOCIAL SECURITY TAX CONTRIBUTIONS
BY COHORT: PRESENT VALUE OF TAX CONTRIBUTION
PER DOLLAR OF PIA

A. Tax Contributions per Dollar of AIME (\$):

Age in 1997	Earnings relative to the average			
	0.5	1.0	1.5	2.0
35	16.8	21.3	22.7	22.0
45	16.4	20.7	22.1	21.4
55	15.1	19.1	20.1	18.9
65	12.1	15.2	15.9	15.6

B. Lifetime Social Security Tax Contribution (Present Value at Age 65 in \$1997):

Age in 1997	Earnings relative to the average			
	0.5	1.0	1.5	2.0
35	150,800	301,600	452,400	592,700
45	139,800	279,500	419,300	546,500
55	122,900	245,800	358,000	419,300
65	84,600	168,300	218,500	243,900

C. Net Cost: Lifetime Taxes Paid Less the Expected Present Worth of Primary Benefits:

Age in 1997	Earnings relative to the average			
	0.5	1.0	1.5	2.0
Women				
35	25,500	103,200	173,300	216,100
45	20,400	90,600	153,400	188,800
55	9,200	65,800	108,900	109,400
65	-13,200	14,000	26,700	25,600
Men				
35	44,700	133,600	216,000	273,900
45	38,700	119,600	194,200	243,700
55	26,600	93,400	147,100	156,900
65	1,700	37,600	56,100	59,100

Notes: See Murphy and Welch (1996) for details of the calculations.

contributions. With a less than fair rate of return, making larger contributions translates into a larger overall loss. This is illustrated in the final panel of Table 2 which gives the net dollar loss (i.e., the present value of taxes paid less the present value of benefits) separately for men and women earning different multiples of the average wage. The lifetime cost to a man currently age 65 who earned twice the average level of earnings over his career is \$59,100, compared to \$273,900 for a man with the same relative earnings history who is 35 years old in 1997. As the table illustrates, the

increase in tax rates and taxable maximums has caused the payoff on Social Security contributions to decline significantly over time.

II. The Basic Problem

Many discussions of the current status of the U.S. Social Security System tend to focus on the prospect of system insolvency. While this prospect is certainly real, focusing on it draws attention away from the more fundamental issue of the large unfunded debt even an actuarially balanced social-security system imposes on today's young. To fix ideas consider the system's budget constraint. The net debt of the system as of date 0 can be written as

$$(1) \quad D_0 = \sum_t 1 / (1 + r)^t B_t - \sum_t 1 / (1 + r)^t T_t - A_0$$

where B_t and T_t are benefits and taxes for year t , r is the rate of return on system assets, and A_0 is the initial level of trust-funds assets. Based on a 3.5-percent real rate of discount and a host of other economic assumptions (see Murphy and Welch, 1996) we estimated the present value of benefits as of 1995 as approximately \$13.7 trillion, while the present value of taxes was about \$10.8 trillion and trust fund assets amounted to only about \$0.4 trillion. This left a net debt of about \$2.5 trillion. According to these calculations, balancing the system would require an increase in OASI taxes of about 23 percent ($=2.5/10.8$) which would raise the OASI component of the Social Security tax rate from its 1993 level of 11.2 percent to 13.8 percent. While precise values for the required tax increase are hard to come by given the large number of assumptions underlying such calculations, 2.6 percentage points is in the range proposed by the Social Security actuaries and others who have examined system solvency.

If we were to raise the OASI tax to 13.8 percent, would Social Security cease to be an issue? It certainly should not. While raising the tax rate to 13.8 percent might bring the overall system into balance, it would only exacerbate the burden the system places on today's young and future generations. To see this, consider a balanced system in which the tax rate has been

increased sufficiently to make the net debt, $D_0 = 0$. We now distinguish between cohorts born at different dates and denote benefits paid to, and taxes collected from, cohort c in year t by B_{ct} and T_{ct} , respectively. For a balanced system, equation (1) implies

$$(2) \quad A_0 = \sum_c \sum_t 1 / (1 + r)^t \{ B_{ct} - T_{ct} \} \\ = \sum_c V_{c0}$$

where

$$V_{c0} = \sum_t 1 / (1 + r)^t \{ B_{ct} - T_{ct} \}.$$

Here, V_{c0} measures the net present value of the net (of taxes) benefits paid to each cohort from date 0 forward. System balance requires that these just offset the system's current assets. While net benefits are positive overall (equal to the value of current assets), they are quite negative for younger cohorts. To illustrate this, we divide the population into two groups, those with positive net benefits (i.e., $V_{c0} > 0$) and those with negative net benefits (i.e., $V_{c0} < 0$). Empirically, the cohorts with positive net benefits correspond to cohorts of men 43 years of age and older and cohorts of women aged 35 and over. The net (of future taxes) payments owed (under current law with the tax rate increased to 13.8 percent) to these older cohorts of men and women amounts to about \$6 trillion. Of this, about half (\$2.9 trillion) is owed to those currently retired. With a trust-fund balance of only \$0.4 trillion the \$6 trillion promised to older cohorts necessitates a net tax on younger cohorts of about \$5.6 trillion (more than twice the value of the Social Security shortfall of \$2.5 trillion and greater than the outstanding national debt). Since the present value of future OASI taxes paid by these younger cohorts is about \$11.2 trillion, the inherited debt accounts for about half of the taxes to be paid by future cohorts. Moreover, these calculations assume that the tax rate is raised immediately to a level required to balance the system. Such changes do not appear to be on the immediate horizon. Any delay in instituting the necessary tax increase would make the magnitude of the

debt imposed on younger cohorts even larger (since older cohorts would pay less of the outstanding obligations).

The low returns on tax contributions for younger cohorts illustrated in the previous section and the \$5.6 trillion debt left for them to pay are really just two sides of the same coin. If the overall system is in balance and debts owed exceed current assets, then the return on contributions for the remaining workers must be low enough to offset this debt exactly.

Such low returns should not be a surprise; low returns and the associated debt are characteristic of a Samuelsonian pay-as-you-go system. As Samuelson pointed out, a steady-state pay-as-you-go system can pay individuals a return equal to the rate of growth of the economy. If the economy grows at 1.5 percent while the rate of discount is 3.5 percent, a simple pay-as-you-go system will pay each cohort only about 56 cents on the dollar, assuming that workers pay taxes over a 40-year career and collect benefits for 15 years after retirement with both taxes and benefits growing with the economy. A lower rate of growth of 0.5 percent (consistent with recent rates of growth for real wages) would imply a return of only 42 cents on the dollar. These numbers are surprisingly similar to what we calculated in the previous section. This suggests that the low returns offered to today's younger workers are largely a necessary by-product of our close to pay-as-you-go system (while not exactly a pay-as-you-go system, the current system is close to one given the low level of system assets and the indexation provisions that cause both taxes and benefits to grow at the growth rate of aggregate earnings). Sadly, it would seem that most of our current problem is really attributable to a problem of design.

Why were earlier cohorts spared such a poor return? The answer is that the pay-as-you-go system was phased in over time by gradually increasing both taxes and benefits to provide higher returns to the early cohorts. While not sustainable, such high returns are always possible in the early stages of a pay-as-you-go system. Today's younger cohorts are really the first to be faced with the low returns inherent in a pay-as-you-go Samuelsonian system.

III. Is Privatization the Answer?

It is tempting to conclude from the above analysis that abandoning the pay-as-you-go system in favor of an alternative funded system with returns tied to the real returns on capital is the natural solution to our troubles. Unfortunately this is not the case. To see this, consider a simple Samuelsonian world with no economic growth where individuals work for one period and then are retired for one period. Starting in year 0, a pay-as-you-go system is instituted, retired workers receive benefits of B , and workers pay taxes of B , so that the system is constantly in balance. With no growth, the system provides workers with a zero rate of return. At some future date, t , we decide to privatize the system. Workers now contribute only the actuarial amount, $C = B/(1 + r)$, required to fund the same B benefit. To keep our commitments to the older cohort in year t , the government must pay out B in period t . It does this by instituting a tax of X per worker starting in year t and borrowing $B - X$ from the private sector. How big must X be? The present value of taxes must be equal to B , so that we must have

$$(3) \quad X(1 + r)/r = B \Rightarrow X = rB/(1 + r).$$

The amount of bonds issued is then $B - X = B/(1 + r)$. What is the total amount of the tax, X , and private pension contribution, C , levied on all cohorts from t forward? The answer is

$$(4) \quad X + C = rB/(1 + r) + B/(1 + r) = B$$

so that every cohort pays exactly the same amount each year as before the change and receives the identical benefit at retirement with absolutely no change in welfare, even though the system is fully privatized. The bond market clears since the private-sector investment of $C = B/(1 + r)$ exactly compensates for the government borrowings in every period. This is true regardless of the level of the private return r and even though the pay-as-you-go example considered had a zero rate of return. Privatization has no effect. Basically, the higher return on the worker's private investment is exactly compensated for by the tax required to pay off the existing liabilities of the

system. This is no coincidence and has nothing to do with the particulars of the example. The basic principle is simple. Since the system balances both before and after the change and since all workers receive the same benefits before and after, the present value of taxes and contributions for the new system must be the same as the present value of taxes before privatization.

One can translate any solvent pay-as-you-go system to an equivalent private system as follows: divide the existing taxes for each worker into two pieces, one piece equal in present value (at the market rate) to his retirement benefit which will be this individual's contribution to his private retirement account and a residual which becomes the new tax (or benefit supplement if it is negative) required to finance the transition. The system balances by construction since all the individual's budget constraints plus the new budget constraint add to the old system-wide budget constraint, and the new individual constraints balance by the definition of providing a fair market return. Moreover, since consumption patterns are unaltered, asset prices need not adjust, and all asset markets will remain in equilibrium.

The above privatization scheme had no effect since it involved no redistributions across individuals or incentive effects. Other privatization schemes will have real effects, but these effects result from redistributions or changes in incentives associated with the method of privatization, and not privatization per se. One could redistribute or change incentives within the pay-as-you-go system and achieve exactly the same results.

What about the difference between the private return and the return paid by the pay-as-you-go system? This simply does not matter. First, the rate of return in a pay-as-you-go system is really a misnomer. There is no actual return since, by definition, nothing is invested in a strict pay-as-you-go system. There are simply a series of transfers, the ratios of which equal the growth rate of the economy. There is no gain to be had since there are no assets to be reinvested at a higher rate of return. The difference in returns between the market and pay-as-you-go has nothing to do with pay-as-you-go per se. The low return paid in the

pay-as-you-go system results entirely from the excess paid to the early cohorts (the initial cohort in the simple Samuelson case). So long as this debt and all obligations to older workers are honored when switching to the private system, privatization provides no gain. The implicit tax of a low return on invested dollars is simply replaced by an explicit tax of an equal amount in present value. The true cost of the pay-as-you-go system is the better than fair deal offered to the early cohorts, not what is done subsequently to pay for these transfers.

IV. What Does Matter?

The arguments presented in the previous section ignored incentives, the heterogeneity of assets, and many of the other aspects of the actual Social Security system. In this section we extend the analysis to consider such complications to see what matters and what does not.

A. Incentives

With flat rate taxes that pay for proportional benefits, the neutrality of the privatization scheme given above would still hold. Essentially, under the pay-as-you-go system the implicit tax rate is simply the difference between the system's tax rate and the actuarially fair contribution, which is the same amount as the constructed tax under the privatization plan. In actuality, however, the Social Security system is far from proportional. The concave benefit structure implies relatively high marginal tax rates. The earnings test and unfair actuarial adjustments provide disincentives to work for those eligible for benefits. A private system with defined contributions, flat tax rates, and actuarially fair benefit schedules would reduce marginal tax rates on work for both the young and old and thereby improve efficiency. Feldstein and Samwick (1997) estimate the efficiency gains from such tax-rate reductions to be substantial (on the order of 2 percent of the tax base). In principle these same gains could be achieved without privatization but may be more difficult to achieve politically.

Another aspect of the current system is that the implicit tax on earnings varies significantly over the life cycle. With a 3.5-percent rate of

discount and a 0.5-percent rate of real wage growth, the value of a dollar contributed to Social Security at age 20 provides only about half of the actuarial return provided by a dollar contributed at age 45 and only one-fourth of the return provided by a dollar contributed close to retirement. Taxes paid in years outside of the "high 35" have absolutely no return. Switching to a defined-contribution approach would improve this aspect as well. It should be noted that such changes in the system will have redistributive effects since the current structure transfers resources to those who retire early, work less after retirement, and have relatively low earnings. Maintaining these transfers to the letter would maintain the current disincentives. However, as Kotlikoff (1996) has pointed out, many of the transfers in the current system (between single and dual income households, for example) seem difficult to rationalize and probably should be eliminated.

B. *Alternative Transition Paths*

The transition path to a privatized system described in the previous section was intentionally constructed in a way that avoided redistribution across cohorts. Given that the current system serves largely as a tax on the young, it may be desirable to shift more of the burden to older cohorts. One way to do this is to balance the system by cutting benefits rather than raising taxes. This would spread the burden across both older and younger cohorts. Alternatively, the debt incurred while privatizing the system can be paid off using a somewhat higher tax rate in the short run and eliminating this transition tax in the long run. While long-run efficiency would be improved by this adjustment, it is not at all clear that this would improve overall efficiency, since the same present value of taxes now will be collected on a smaller base (usually leading to greater overall distortion). Moreover, such a transition would concentrate the burden of the \$5.6 trillion overhang on the transition cohorts rather than on all future taxpayers.

C. *Changing the Tax Base*

So far we have considered programs that fund the Social Security taxes from the tax

base of labor earnings such as the current Social Security tax. As Kotlikoff (1996) points out, switching to an alternative tax base for the transition tax (such as consumption) may generate significant efficiency gains. Clearly, such gains in principle could be generated without changing to a privatized system.

D. *Differences in the Rates of Return Across Assets*

Currently, trust-fund assets are invested in special U.S. Treasury securities. Some have suggested that it would be advantageous to shift these holdings to higher-yield assets such as broad-based mutual funds either by privatization or by the trust fund itself investing in private equities or bonds. First, the issue of the types of assets held is of almost no relevance for most of the historical period since the system has maintained very little in assets for most of its history. Second, if there is a gain to be had by switching to holding private assets then such a gain is independent of even the existence of Social Security. This can be seen by comparing two scenarios. In the first scenario, the trust fund buys fewer government bonds and instead buys private-sector assets. This requires the private sector to buy the extra bonds and sell the private-sector assets. This is exactly the same series of transactions required if the government were simply to sell more bonds and use the proceeds to purchase private-sector assets without regard to Social Security or the trust fund. If there is a gain to such actions, then the government has a money machine that it is not taking full advantage of, regardless of the fate of the Social Security trust fund. Such a money machine might exist, for example, if government debt is valued for its liquidity. In such cases this debt can earn a lower than market rate of return and be a source of revenue for the government similar to the government revenues from money creation.

In general, if markets are complete then transactions of the type described above will have no real effects provided that they do not change government consumption and private wealth holdings are perfect substitutes for the implicit holding through government assets (i.e., a Ricardian world).

Private-sector transactions will just undo whatever the government does so as to maintain the same total holdings at the same asset prices. To the extent that privately held assets and government obligations are not perfectly substitutable or markets are not complete then changing the investment portfolio might matter (i.e., those who hold implicit obligations to or claims on the government might prefer a more risky [or less risky] public portfolio). Such arguments are hard to evaluate, however, since the identity of the marginal claimants on government returns is difficult to know. Privatization goes some distance toward solving this problem by giving individuals greater control over their portfolios rather than relying on their ability to compensate for any undesirable aspect of the government investment policy.

E. Political Risk

One very undesirable feature of the current system is the uncertainty over what will be done to make it solvent. Many workers, particularly younger workers, have very little idea of what will be available for them through the Social Security system. One major problem with even a fixed up pay-as-you-go system is that benefit levels and taxes are always subject to change based on political forces. A defined-contribution system would eliminate much of this political risk while substituting perhaps more economic risk.

F. Political Commitment

The current problems with Social Security stem mostly from the better than actuarially fair deals given to past workers. Our reluctance to pay for these transfers by allowing the system to grow relative to the economy for the past five decades only made things worse. Most likely, such large transfers would have been impossible had they been proposed on their own (on top of, say, a funded actuarially fair system for younger workers). Moreover, it is always tempting to continue to make such transfers in the future. Preventing ourselves from digging an even deeper hole is probably one of the most com-

pellent reasons for getting away from a pay-as-you-go structure.

G. Doing Something Sooner Rather Than Later

Doing something sooner rather than later is probably the most important thing. The longer we wait to make the system solvent and reduce the burden on future generations, the more difficult it will be to do. The costs of waiting are easily calculated. From equation (1) we can write the deficit as

$$(5) \quad D_0 = \sum 1 / (1 + r)^t B_t - \sum 1 / (1 + r)^t T_t - A_0.$$

This implies that the percentage increase in taxes required from date t forward to balance the system is

$$(6) \quad \Phi_t = D_0 / \left[\sum_{\tau=t}^{\infty} 1 / (1 + r)^{\tau} T_{\tau} \right] \\ \Rightarrow (\Phi_{t+1} - \Phi_t) / \Phi_t \\ = T_t / \left[\sum_{\tau=t}^{\infty} 1 / (1 + r)^{\tau-t} T_{\tau} \right]$$

so that the required tax increase grows at a rate equal to the ratio of the current taxes to the present value of all future taxes. For an economy growing at rate g with an interest rate of r , the required tax rate would grow at the rate $r - g$. Hence if balancing the system would require an immediate 2.6-percentage-point increase in the tax rate, the required tax increase would be 3.5 percentage points in ten years and 4.7 percentage points in 20 years, based on a 3.5-percent real rate and 0.5 percent real growth. The same base calculations hold for changes in benefits. Clearly, the longer we wait, the greater will be the required adjustments, and the more of the burden we will pass on to future generations.

REFERENCES

- Diamond, Peter. "The Future of Social Security," in Peter Diamond, David Linderman,

- and Howard Young, eds., *Social Security: What role for the future?* Washington, DC: National Academy of Social Insurance, 1996, pp. 225-33.
- Feldstein, Martin and Samwick, Andrew. "The Transition Path in Privatizing Social Security." National Bureau of Economic Research (Cambridge, MA) Working Paper No. 5761, 1996.
- _____. "The Economics of Prefunding Social Security and Medicare Benefits." National Bureau of Economic Research (Cambridge, MA) Working Paper No. 6055, 1997.
- Kotlikoff, Laurence. "Privatizing Social Security: How It Works and Why It Matters," in James Poterba, ed., *Tax policy and the economy*, Vol. 10. Cambridge, MA: MIT Press, 1996, pp. 1-32.
- _____. "Simulating the Privatization of Social Security in General Equilibrium," in Martin Feldstein, ed. *Privatizing Social Security*. Chicago: University of Chicago Press, 1998 (forthcoming).
- Murphy, Kevin M. and Welch, Finis. "Real Wage Growth and OASDI Trust Fund Solvency." Working paper, University of Chicago, 1996.

Social Security and the Real Economy: An Inquiry into Some Neglected Issues

By ISAAC EHRLICH AND JIAN-GUO ZHONG*

The debate concerning the effects of pay-as-you-go (PAYG), defined-benefits, social-security systems on the real economy has focused on private savings (Robert J. Barro, 1978; Martin Feldstein, 1997). We expand the inquiry to neglected effects on economic growth and underlying family choices. Our inquiry is based on Ehrlich and Francis T. Lui's (1998) model of the relationships among social security, the family, and endogenous growth.

I. The Family's Role in Economic Growth

We view the family, in part, as an efficient partnership across overlapping generations. Altruism is a major motivating force for parents, but we also consider a complementary incentive: old-age insurance. The family's overlapping generations are linked by mutual dependency. Children rely on their parents for nurture and investment in productive knowledge, and retired parents rely on their children for emotional and material support, including informal care. The family has thus traditionally served as an informal self-insurance, or "family-security," setup. This insurance differs from market insurance or savings because it involves productive intergenerational transfers.

An opportunity for such transfers exists because persistent human-capital accumulation, hence growth, must rely, at least in part, on inputs by the older generation into the education of the young. A simple specification of the production technology is given by

$$(1) \quad H_{t+1} = A(\bar{H} + H_t)h_t$$

where H_{t+1} and H_t denote the attained human-capital stocks of child and parent(s), respectively, \bar{H} denotes raw labor, h_t is the fraction of the production capacity parents choose to devote to the education of each child, and A is a technological parameter.

The family partnership can take advantage of this opportunity through an efficient sharing arrangement, supported by implicit contracts: children share in the material success of their parents because parents' investments in them are proportional to the parents' production capacity. Old parents can also share in the success of their grown-up children in some proportion to their contribution to the latter's quantity and human capital. When determined efficiently, this sharing arrangement turns out to be an important force, complementary to altruism, in promoting human-capital accumulation and growth. The incentive to uphold this insurance system exists regardless of whether some old-age needs can also be secured through savings or private pensions.

A major concern about family insurance is the risk of default because of nonsurvival or noncompliance of children with implicit family contracts, especially when family size declines. But our basic predictions hold even if we allow for such default risks. Moreover, they hold even if family insurance is inoperative, and the only force linking parents and children is parental altruism (see Ehrlich and Lui, 1991, 1998).

II. The Effect of Social Security on Family Choices

Our central proposition is that an increase in a social-security tax that is proportional to earnings will affect adversely at least one of three variables controlling the economy's growth path: fertility, investment in human capital, or the savings rate (and possibly all three). Moreover, if the tax reduces the incentive to bear or invest in children, it indirectly

* Department of Economics, State University of New York, Buffalo, NY 14260, and PNC National Bank, 300 Bellevue Parkway, Wilmington, DE 19809, respectively. We are indebted to Jinyoung Kim, Yong Yin, and Ted Bergstrom for valuable comments and suggestions.

reduces the gains from marriage (Gary S. Becker, 1991). The PAYG system can thus be expected to affect adversely the incentive to form families.

The rationale for this proposition is as follows: the source of the predicted adverse effects of the social-security tax is an externality or "moral hazard." While in most families the tax is levied in proportion to earnings, the benefits received by an old parent are "defined"; that is, they are independent of what the parent has actually contributed, and certainly of what the parent's own children contribute to the social fund. There is thus no incentive to take the latter into account when investing in children.

More formally, an increase in the social-security tax rate reduces the rate of return to parents from children, relative to their intertemporal rate of substitution in consumption. This is true even if parents are motivated purely by altruism, or if family insurance is subject to default risks. Savings are also affected because, in equilibria involving interior solutions, the rates of return from savings and all other investments are equated on the margin. Families subject to a lump-sum tax (if their earnings exceed the maximum taxable level), however, would generally be immune to the adverse consequences of the tax because it may not interfere with their marginal investment or sharing decisions.

The specific effects of the "proportional" tax on each of the three determinants of the economy's growth path will vary over different stages of economic development. At an early phase, where fertility is high and investment in human capital is low, fertility is more likely to be adversely affected. The effects on savings and human-capital investment are ambiguous, but they must go in the same direction. At an advanced stage of development, where fertility may fall to its lower bound (consistent with the stylized facts of the "demographic transition"), the tax's effects on fertility would be small. Its main "casualty" in a growth equilibrium is likely to be savings and investment in human capital, hence economic growth.

Alternative stages of development are themselves an outcome of the model's basic parameters: the technology of learning and the probabilities of survival of children to adult-

hood and of adults to old age, π_1 and π_2 . If the latter are sufficiently low, the economy will be stuck at a "stagnant equilibrium" involving a high fertility and a low level of investment in human capital. A sufficient upward shock in any one of these variables, however, can induce a takeoff from a stagnant equilibrium to persistent growth equilibrium, and an associated "demographic transition." Moreover, the steady-state growth rate itself is an increasing function of π_1 , although not necessarily of π_2 if the latter affects mainly the incentive to save. The latter is generally a decreasing function of π_1 .

III. Empirical Implementation

We test these propositions against international panel data including a measure of the social-security tax rate. While our preceding analysis treats the latter as exogenous, we allow for its possible endogeneity. We account for country-specific "missing" factors via fixed-effects regressions, which separate within- and between-country variations in key regressors.

A. The Sample

Data availability limit our sample to 49 countries over 29 years: 1960–1989. For some countries, data are complete for shorter periods. This produces 673 observations.

We measure our theoretical tax rate by the "pension" portion of social-security benefits relative to GDP (PEN), as reported by the International Labor Office.¹ Real per capita income (GDP) and the GDP shares of government spending (G) and private investment (I) are from Robert Summers and Alan Heston (1988). The shares of government deficit (DEFICIT) and current account surplus (NX) in GDP, taken from IMF statistics, along with I enable us to approximate the average savings rate (S). The total fertility rate (TFR) and

¹ We select only the "pension" portion of social-security payments, which includes old age, disability, and survivor benefits as a percentage of GDP. Since in three countries PEN is zero, in the log transformations we replace 0 by 0.00001, a value much below the minimal PEN.

the marriage (MARRY) and divorce (DIVORCE) rates, as well as the survival probabilities to adulthood and old age (Pi1 and Pi2), AGE (0-14), and SEX (female/male ratio) are from the *Demographic Yearbook* of the United Nations.

Since we expect some of the effects of PEN to vary by the country's stage of development, we account for the latter partly by separating our "full" sample to "rich" and "poor" subsets. To achieve a balanced split, we distinguish countries above and below a per capita GDP level of \$7,650 in 1980 (the average for our sample).²

B. The Regression Model

The basic specification is a multiequation log-linear, fixed-effects model:

$$(2) \quad Ly = \alpha_0 + \alpha_1 LPEN + \alpha_2 LPi1 + \alpha_3 LPi2 + \alpha_4 LG + \alpha_5 LGDP$$

where $y = TFR, S, MARRY$, and $DIVORCE$; α_0 represents country-specific effects; and L denotes natural logs. The regressors in (2) are the measurable parameters of the model, supplemented by relevant initial conditions. We test for the endogeneity of PEN using Hausman's test. Where relevant, we apply a two-stage least-squares analysis to test the effect of PEN.

Pi1 and Pi2 (proxies for π_1 and π_2) are measured as the survival probabilities of the population (based on age-specific death rates) from ages 0 to 24 and 50 to 74, respectively. Since the left-hand-side variables approximate steady-state values of the endogenous variables of the model along the equilibrium path,

we also enter current GDP to account for transitional deviations of the actual variables from their long-run equilibrium values. The share of government spending in GDP, G , is introduced in order to isolate the effect of PEN from that of other government spending, which may exert an independent influence on the growth path's determinants.

A special econometric specification is applied to regressions concerning the long-term growth rate (g). In a growth equilibrium,

$$(3) \quad GDP_t = (GDP_0) \exp[g(X)t]$$

and by the logic of our model:

$$(4) \quad g(X) = \beta_1 + \beta_2 LPEN + \beta_3 LPi1 + \beta_4 LPi2 + \beta_5 LG.$$

Taking the log of (3), the growth rate of equation (4) can then be estimated from:

$$(5) \quad LGDP = \beta_0 + \beta_1 t + \beta_2 t LPEN + \beta_3 t LPi1 + \beta_4 t LPi2 + \beta_5 t LG + \beta_7 LPEN + \beta_8 LPi1 + \beta_9 LPi2 + \beta_5 LG$$

where β_0 represents country-specific effects. The combined effect of terms involving the time trend t , (T), is taken to be an unbiased estimate of g in the separate country sets. The interaction terms of t with the variables represented by X account for the effect of both between- and within-country variations in X on g . The levels of X account for short-run level effects of within-country variations in X on GDP.

IV. Analysis of the Results

A. Growth Regressions

In Table 1 we estimate equation (5). Model 1 includes all the regressors in log form and treats PEN as exogenous. We also ran model 1 with PEN entered in its linear form simply as a robustness test (see footnote 1), but the results were comparable.

² The "poor" set includes the Philippines, El Salvador, Thailand, Jamaica, Tunisia, Korea, Costa Rica, Jordan, Panama, Fiji, Czechoslovakia, Malaysia, Chile, Poland, Mauritania, Brazil, Portugal, Hungary, Yugoslavia, Greece, Mexico, Argentina, Ireland, Singapore, Spain, Venezuela, and East Germany. The "rich" set includes Israel, Hong Kong, Japan, the United Kingdom, Italy, New Zealand, Australia, Finland, Belgium, Trinidad, Netherlands, Denmark, France, Luxembourg, West Germany, Iceland, Norway, Sweden, Austria, Canada, Switzerland, and the United States.

TABLE 1—GROWTH REGRESSIONS
(DEPENDENT VARIABLE = LGDP)

A. Model 1			
Independent variable	Full set	Rich set	Poor set
T	0.0289 (4.887)	-0.0017 (-0.316)	0.0591 (5.011)
TLPEN	-0.0024 (-7.814)	-0.0044 (-7.395)	-0.0013 (-3.127)
TLPi1	0.0844 (2.599)	-0.2162 (-3.853)	0.2045 (4.822)
TLPi2	-0.0168 (-4.317)	-0.0009 (-0.325)	-0.0400 (-4.546)
TLG	-0.0052 (-2.855)	0.0026 (1.699)	-0.0156 (-4.833)
LPEN	0.0373 (2.736)	0.1394 (8.744)	-0.0030 (-0.155)
LPi1	-0.4651 (-0.970)	2.6943 (4.135)	-1.6791 (-2.338)
LPi2	0.2806 (3.332)	0.0139 (0.229)	0.2741 (1.386)
LG	-0.3679 (-7.484)	-0.2695 (-5.677)	-0.2674 (-3.165)
Adjusted R ² :	0.9652	0.9347	0.9283
N:	673	359	314

B. Model 2

Independent variable	Full set	Rich set	Poor set
T	0.0306 (3.826)	-0.0015 (-0.325)	0.0726 (4.665)
TELPEN	-0.0024 (-7.729)	-0.0034 (-6.552)	-0.0012 (-2.916)
TLPi1	0.0927 (2.746)	-0.2369 (-4.881)	0.2077 (4.888)
TLPi2	-0.0169 (-2.985)	-0.0037 (-1.548)	-0.0320 (-3.013)
TLG	-0.0056 (-3.071)	0.0008 (0.584)	-0.0178 (-4.911)
ELPEN	0.0218 (0.194)	0.2926 (14.373)	-0.1317 (-1.373)
LPi1	-0.4821 (-0.917)	2.8782 (5.252)	-1.3268 (-1.747)
LPi2	0.2879 (2.579)	0.0619 (1.221)	0.0868 (0.355)
LG	-0.3308 (-6.508)	0.0212 (0.452)	-0.1399 (-1.109)
Adjusted R ² :	0.9655	0.9541	0.9286
N:	670	356	314

Notes: Rows show the estimated β and β/SE_{β} (in parentheses). Results for country-specific dummies are suppressed. Model 2 treats LPEN as endogenous via a two-stage least-squares analysis.

Hausman's test indicates that LPEN is an endogenous variable in model 1. In model 2, LPEN is therefore replaced by its expected

value, ELPEN, predicted from a reduced-form regression containing the exogenous variables in model 2 plus AGE (0–14), SEX, and the initial value of LPEN in any country. A complete analysis of the determinants of PEN is beyond our scope, although it is reasonable to expect that public support for a PAYG social-security system would be greater in countries with small families, aging populations, and higher incomes. This is borne out by the reduced-form results.

Table 1 strongly supports our predictions. The pension variable has an adverse effect on the growth rate, regardless of the influence of G or any other corrections, but especially so in the rich-countries set. The effect of PEN in the set of poor countries is significantly smaller or insignificant. The survival probabilities have their generally predicted effects on the growth rate, although in the rich set Pi1 has the wrong sign. Government spending, G , generally shows an adverse effect on growth.

B. Fertility, Marriage, and
Divorce Regressions

The dependent variables in Table 2A are the number of children born to an average female aged 15–49 (TFR; a proxy for the number of children per family) and the annual rates of marriages and divorces in the population, all in log form. The regressions implement equation (2), supplemented by relevant demographic variables. Also, based on Hausman's tests we reject the exogeneity of LPEN in the fertility and divorce regressions, and the effect of LPEN in these regressions is estimated via the two-stage least-squares procedure. The basic results are again supportive of our basic hypotheses. The pension variable has a negative effect on fertility, primarily in the poor set of countries. In the rich set, the effect of pension is smaller and sometimes insignificant statistically. Note that PEN does not include aid to families with dependent children, which can be expected to raise the incentive to bear children. Since the "missing" AFDC variable is likely to be positively correlated with PEN, the latter's adverse effects on fertility may be understated.

In the marriage and divorce regressions, the pension variable has our predicted adverse effect on family formation in all countries. In-

TABLE 2—FERTILITY, FAMILY FORMATION,
AND SAVINGS REGRESSIONS

A. Fertility and Family Formation				
Independent variable	Fertility ^a		Family formation (full set)	
	Rich set	Poor set	Marriage ^b	Divorce ^c
LPEN	-0.0037 (-0.201)	-1.4864 (-2.852)	-0.1144 (-6.094)	1.5169 (5.139)
LPi1	-3.5664 (-5.761)	-3.3511 (-5.513)	1.1552 (1.86)	-11.1146 (-3.715)
LPi2	-0.0605 (-1.883)	-0.0580 (-1.828)	-0.1073 (-2.348)	-0.1822 (-2.586)
LGDP	-0.4733 (-11.357)	-0.4209 (-9.531)	0.0620 (1.277)	0.5042 (6.603)
LG	-0.2334 (-3.348)	-0.2082 (-3.006)	0.0400 (0.600)	0.8590 (8.666)
LMARRY	0.0778 (2.201)	0.0536 (1.495)		
LAGE			0.2853 (2.749)	-0.8887 (-4.345)
Adjusted R ² :	0.9041	0.9063	0.4135	0.9196
N:	391	391	641	529
B. Savings				
Independent variable	Savings (full set)			
	Model 1 ^d	Model 2 ^e		
LPEN	-95.679 -8.319	-129.4685 -6.932		
LPi1	4.7288 0.384	-63.3755 -3.169		
LPi2	-12.2927 -4.366	-9.6485 -2.076		
LGDP	-0.6092 -6.06	-1.5243 -9.724		
LG	0.0003 1.884	0.0002 0.721		
LMARRY				
LAGE	-40.8669 -5.338	-93.4227 -7.652		
NX	-0.2970 -7.421			
DEFICIT	-0.0172 -0.329			
Adjusted R ² :	0.7832	0.8028		
N:	601	601		

Notes: Rows show the estimated β and β/SD (in parentheses). Results for country-specific dummies are suppressed. In both the fertility and divorce models, LPEN is treated as endogenous via a two-stage least-squares analysis. The savings regressions are in linear form. Model 1 is unrestricted; model 2 is a restricted regression, forcing a coefficient of -1 to NX and DEFICIT.

^a Dependent variable = LTFR.

^b Dependent variable = LMARRY.

^c Dependent variable = LDIVORCE.

^d Dependent variable = I .

^e Dependent variable = $S = I + \text{DEFICIT} + \text{NX}$.

our “tax” measure (PEN). Note that in some countries, it is possible that the social-security provisions provide fuller pension benefits for homemakers who stay married to a spouse meeting the minimal labor-market requirements, which would generate an offsetting influence on the incentive to marry and divorce. Nevertheless, PEN exerts a significant inducement on the margin for couples to opt out of marriage in all countries.

C. Savings Regressions

We do not have private savings data for our full sample. We resort, instead, to the investment share of GDP (I) as a proxy for our theoretical savings rate, S . In countries subject to large variations in government deficits (DEFICIT) or current account surpluses (NX), however, (I) would be an inaccurate proxy for S . In model 1 of the savings regressions of Table 2B, we therefore introduce DEFICIT and NX as additional regressors. This model provides an unrestricted estimate of the effect of PEN on the propensity to save. We are forced to use a linear-regression format in this model, because DEFICIT and NX often involve both positive and negative values.

In model 2, we use the national income identity $S = I + \text{DEFICIT} + \text{NX}$ to create a direct “savings” measure. This restricted regression forces the coefficients of DEFICIT and NX to equal -1. Hausman’s test cannot reject the hypothesis that PEN is exogenous in the context of the “savings” regressions, and therefore no two-stage least-squares analysis is pursued. The results of both the restricted and unrestricted regressions indicate that PEN has a significant depressing effect on savings, and the restricted estimates actually exhibit a greater explanatory power. Also, the propensity to save is inversely related to Pi1 in the restricted regressions. However, our expectation that it would be directly related to Pi2 is not supported by this analysis.³

deed, it significantly decreases the marginal propensity for marriage while increasing the marginal propensity for divorce. These results support both our model and the relevance of

³ To test the robustness of our results in Table 1 we added interaction terms involving T and lagged GDP, initial GDP, or AGE, and we introduced all regressors in their initial values to allow for the endogeneity of Pi1 and Pi2, as predicted by Ehrlich and Hiroyuki Chuma (1990),

V. Some Policy Implications

The estimated effects of PEN in Tables 1 and 2 are not just significant statistically, but also quantitatively. In the growth regressions based on model 2 of Table 1 for the full data set, a 1-percentage-point increase in PEN from its mean of 0.0593 would induce a reduction of -0.0004 percentage points from the annual growth rate of countries in that set. Had the average value of PEN remained constant at its 1960 fraction of GDP (0.0385), this would have increased the annual growth rate over the sample period from 2.74 percent to 2.82 percent. The corresponding analysis for the rich country set shows a potential increase in the growth rate from 2.35 percent to 2.47 percent. In the United States the potential increase in the growth rate from 2.1 percent to 2.21 percent would have added 3 percent to per capita GDP in 1989.

Similarly, we calculate that a 1-percentage-point rise in PEN will induce a reduction of 0.22 in TFR from its mean of 3.1 in the poor country set (the effect in the rich set is insignificant). In the full data set we expect the same rise in PEN to decrease the rate of marriage by 0.1 percent from its mean of 7 percent; increase the rate of divorce by 0.3 percent from its mean of 1.4 percent; and lower the savings rate in GDP by 1.3 percent from its mean of 18.9 percent. Of course, all these quantitative effects must be viewed with considerable caution because of our imperfect data.

Whatever limitations exist in our analysis, however, taken together the results appear to be consistent across the various regressions and are generally strongly supportive of our

specific theoretical propositions. They imply that, in the debate about the merits of our current social-security system, we should consider the effects this system may have on the real economy. Many Western countries face the specter of a financial collapse of their systems early in the next century because of a continuous aging of the population and a slowdown in productivity growth. Our analysis implies that, in addition to adverse effects on private savings, the PAYG social-security systems may exert significant adverse effects on the rate of productivity growth, especially in the more developed countries, and on fertility, especially in the developing countries, which in turn would exacerbate the financial problem.

Reforming Social Security into a fully funded and privately managed system, which would avoid the major disincentive effects of the PAYG system, may thus be a good recipe not just for prudent fiscal policy, but also for achieving a more efficient, pro-growth, and pro-family economic policy.

REFERENCES

- Barro, Robert J. *The impact of Social Security on private savings: Evidence from the US time series*. Washington, DC: American Enterprise Institute, 1978.
- Becker, Gary S. *A treatise on the family*. Cambridge, MA: Harvard University Press, 1991.
- Ehrlich, Isaac and Chuma, Hiroyuki. "A Model of the Demand for Longevity and the Value of Life Extension." *Journal of Political Economy*, August 1990, 98(4), pp. 761-82.
- Ehrlich, Isaac and Lui, Francis T. "Intergenerational Trade, Longevity, and Economic Growth." *Journal of Political Economy*, October 1991, 99(5), pp. 623-48.
- _____. "Social Security, the Family, and Economic Growth." *Economic Inquiry*, 1998 (forthcoming).
- Feldstein, Martin. "Social Security and Saving: New Time Series Evidence." *National Tax Journal*, June 1996, 49(2), pp. 151-64.
- International Labor Office. *The cost of social security*. Geneva, Switzerland: International Labor Office, 1997.

which improved the results. We tested and corrected for serial correlation. We introduced a time trend in the regressions of Table 2. We also conducted Box-Cox analyses of optimal transformations. Our main results are robust to all tests. Finally, we excluded from our sample Singapore and Malaysia, where social security is a provident fund, rather than a strict PAYG system, since by our analysis, the former would be immune to the predicted adverse incentive effects of the latter on family choices. Exclusion of these countries from our sample, results in even sharper support for our predicted effects.

International Monetary Fund. *International financial statistics*. Washington, DC: International Monetary Fund, various years.

Summers, Robert and Heston, Alan. "A New Set of International Comparisons of Real Prod-

uct and Price Levels Estimates for 130 Countries." *Review of Income and Wealth*, March 1988, 34(1), pp. 1-25.

United Nations. *Demographic yearbook*. New York: United Nations, various years.

SOCIAL SECURITY AND DECLINING LABOR-FORCE PARTICIPATION: HERE AND ABROAD[†]

Social Security and Retirement: An International Comparison

By JONATHAN GRUBER AND DAVID WISE*

In almost every industrialized country, the population is aging rapidly, and individuals are living longer. The ratio of the number of persons age 65 and over to the number age 20–64 is shown Figure 1 now and in future years for 11 countries. The increase is striking in almost every country. In Japan, with the most rapid population aging, the ratio will more than double by 2020 and will almost triple by 2050. These demographic trends have placed enormous pressure on the financial viability of the social-security systems in these countries. The financial pressure caused by demographic trends is compounded by another trend. In virtually every country, employees are leaving the labor force at younger and younger ages. The trend is most evident for men of course, but for older women participation is also de-

clining, in spite of large increases in the labor-force participation of younger women. In some countries, the labor-force participation rates of 60–64-year-old men have fallen by 75 percent over the past three decades, increasing substantially the proportion of retired persons to those in the labor force.

One explanation for the striking decline in labor-force participation is that social-security provisions themselves provide enormous incentive to leave the labor force early, thus by their very structure exacerbating the financial problems that they face. This is the aspect of social-security plan provisions that is emphasized in this paper. By considering the relationship between plan provisions on the one hand and labor-force participation rates on the other, we draw attention to the important role that social security can have on the labor-force decisions of older persons.

This paper presents comparisons based on evidence presented in papers written for 11 industrialized countries. The authors, who are listed below, in the * footnote, are part of an ongoing project to analyze the relationship between social-security plan provisions and retirement in many countries, as well as other related issues. Three of the individual country papers are summarized by other participants in this session. Important conclusions that can be drawn from comparison of the findings of all of the individual papers are distilled in this brief paper. The project relies on the analysis of a large group of economists who have analyzed social-security provisions and labor-force participation in their own countries. The central feature of the project is the presentation of comparable descriptive data and analytic calculations for each of these 11 countries. Thus, comparisons can be made on a common

[†] *Discussants:* John Rust, Yale University; Andrew Samwick, Dartmouth College.

* Gruber: Department of Economics, Massachusetts Institute of Technology, Cambridge, MA 02139, and NBER; Wise: J. F. Kennedy School of Government, Harvard University, 79 John F. Kennedy St., Cambridge, MA 02138, and NBER. This work is based on papers by Pierre Pestieau and Jean-Philippe Stijns (1998), Belgium; Jonathan Gruber (1998), Canada; Didier Blanchet and Louis-Paul Pelé (1998), France; Axel Börsch-Supan and Reinhold Schnabel (1998b), Germany; Agar Brugiavini (1998), Italy; Takashi Oshio and Naohiro Yashiro (1998), Japan; Arie Kapteyn and Klaas de Vos (1998), Netherlands; Michele Boldrin, Sergi Jimenez-Martin, and Franco Peracchi (1998), Spain; Mårten Palme and Ingemar Svensson (1998), Sweden; Richard Blundell and Paul Johnson (1998), United Kingdom; and Peter Diamond and Jonathan Gruber (1998), United States. The work was supported by the National Institute on Aging through grant no. 5 P20 AG12810 to the National Bureau of Economic Research. We are indebted to comments from all of the country-paper authors in putting together this comparison.

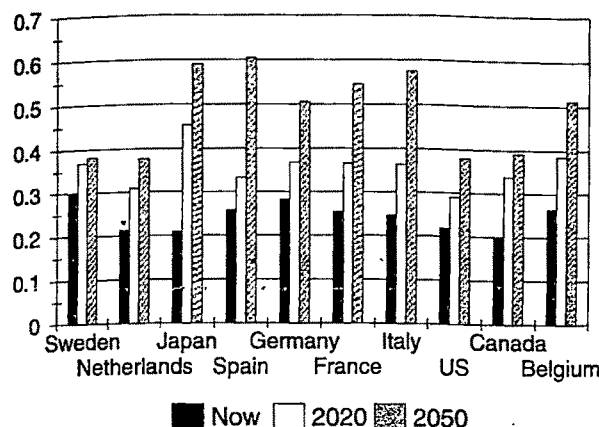


FIGURE 1. RATIO OF POPULATION AGE 65+ TO POPULATION AGE 20-64

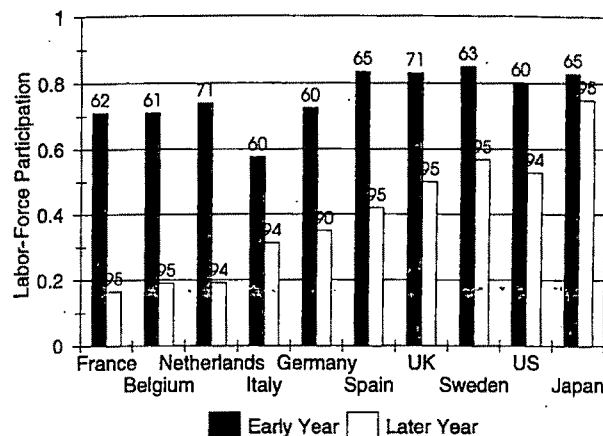


FIGURE 2. DECLINE IN LABOR-FORCE PARTICIPATION, AGE 60-64

Note: Numbers above histogram bars give years corresponding to the data for each country.

footing. The core of each country paper is a detailed analysis of the retirement incentives inherent in the provisions of that country's retirement-income system. By making the same analytic calculations, the individual studies provide a means of comparing the retirement incentives among the nations. To facilitate comparisons across countries, we refer to data for men in this paper. The individual country papers present parallel data for both men and women, and it is clear that the incentive effects of social-security plan provisions are important for women as well as for men.

I. Decline in Labor-Force Participation

The decline in the labor-force participation of older persons is perhaps the most dramatic feature of labor-force change over the past several decades. The decline has been striking in all but one of the countries studied here. The labor-force participation rates of men aged 60-64 for the years 1960-1996 are shown for each of the 11 countries in Figure 2. The decline was substantial in each of the countries but was much greater in some countries than in others. In the early 1960's, the participation rates were above 70 percent in all but one of the countries and above 80 percent in several countries. By the mid 1990's, the rate had fallen to below 20 percent in Belgium, Italy, France, and the Netherlands. It had fallen to about 35 percent in Germany and 40 percent in Spain. Although analysts in the United

States have often emphasized the "dramatic" fall in that country, the U.S. decline from 82 percent to 53 percent was modest in comparison to the much more precipitous decline in these European countries. The decline to 57 percent in Sweden was also large, but modest when compared to the fall in other countries. Japan stands out with the smallest decline of all the countries, from about 83 percent to 75 percent. Labor-force participation rates of 45-59-year-old men, as well as those 60 and older, have also declined substantially, and these trends can be seen in the individual country papers.

The current relationship between labor-force participation and age for men is shown for each of the countries in Figure 3. At age 50, approximately 90 percent of men are in the labor force in all of the countries. The decline after age 50 varies greatly among countries. By age 65 fewer than 5 percent of men in Belgium are working, and in all but three countries fewer than 20 percent are working. The range in participation rates is large, however. In Japan almost 75 percent of men are still in the labor force at age 60, and 60 percent are still working at age 65.

There are many implications of the withdrawal of older men from the workforce. We emphasize the forgone productive capacity of older employees who leave the workforce. Here we use a rather crude measure of forgone

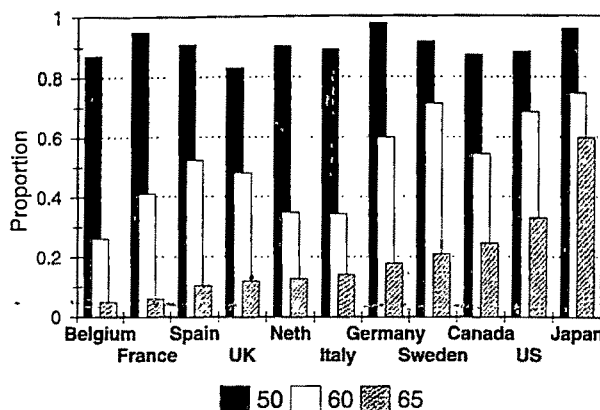


FIGURE 3. CURRENT LABOR-FORCE PARTICIPATION BY COUNTRY AND AGE

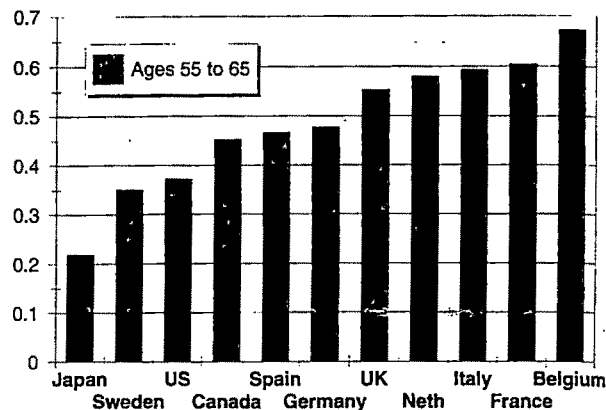


FIGURE 4. UNUSED PRODUCTIVE CAPACITY (PROPORTION)

productive capacity. Consider the proportion of men not working at a given age ($1 - \text{LFP}$, where LFP is the labor-force participation rate): about 0.95 in Belgium and about 0.40 in Japan at age 65, for example. Loosely speaking, we refer to this measure as the “unused productive capacity” at that age. If the unused capacity is added up over all ages in some range, we find the area above the LFP curve in that range. When divided by the total area above and below the curve for that age interval and multiplied by 100, it provides a rough measure of the unused capacity over the age interval, as a percentage of the total labor capacity in that age range.

The unused productive capacity measures for all of the countries are shown in Figure 4 for the 55–65 age group. Unused capacity in this age group ranges from 67 percent in Belgium to 22 percent in Japan. We consider below how this relative measure is related to the provisions of the social-security programs in the countries.

II. Social-Security Benefit Accrual and the Implicit Tax on Work

Two features of social-security plans have an important effect on labor-force participation incentives. The first is the age at which benefits are first available. This is called the early-retirement age. The “normal” retirement age is also important but is typically much less important than the early-retirement age. It may once have been that the normal

retirement age was when most people were expected to retire; now in most countries, few people work until the “normal” retirement age.

The extent to which people continue to work after the early-retirement age is closely related to the second important feature of plan provisions, the pattern of benefit accrual. Suppose that at a given age a person has acquired entitlement to future benefits upon retirement. The present discounted value of these benefits, minus future taxes paid, is the person’s social-security wealth at that age (SSW_a). The key consideration for retirement decisions is how this wealth will evolve with continued work. If a person is 59, for example, what is the change in SSW if he retires at age 60 instead of age 59? The difference between SSW if retirement is at age a and SSW if retirement is at age $a + 1$, $\text{SSW}_{a+1} - \text{SSW}_a$, is called SSW accrual.

We compare the SSW accrual to net wage earnings over the year. If the accrual is positive, it adds to total compensation from working the additional year; if the accrual is negative, it reduces total compensation. The ratio of the accrual to net wage earnings is an implicit tax on earnings if the accrual is negative, and an implicit subsidy to earnings if the accrual is positive. Thus a negative accrual discourages continuation in the labor force, and a positive accrual encourages continued labor-force participation. This accrual rate (along with the associated tax rate) is a key calculation that is made in the same way for

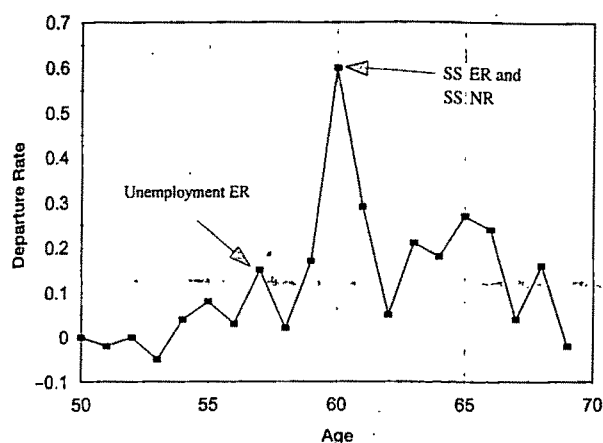


FIGURE 5. HAZARD RATES FOR FRANCE

Notes: ER = early retirement; NR = normal retirement. Persons who become unemployed between ages 57 and 59 are guaranteed income approximately equal to social-security benefits which will begin at age 60.

Source: Blanchet and Pelé (1998).

each of the countries considered here. As it turns out, the pension accrual is typically negative at older ages: continuation in the labor force implies a reduction in the present discounted value of pension benefits. That is, in most countries, due to insufficient actuarial adjustment for fewer years of pension receipt, combined with generous earnings replacement rates for retirees and high social security payroll taxes for workers, there is an implicit tax on work and an incentive to leave the labor force. The magnitude of the SSW accrual and the corresponding tax or subsidy differ greatly from country to country.

Two features of plan provision are particularly important: the age the benefits are first available and the tax on earnings if a person continues to work after this age. Each is discussed in turn.

III. The Importance of the Early-Retirement Age

For illustration, we draw here on data from France. As in all the countries, the current labor-force departure rates in France correspond closely to social-security provisions. Social-security benefits in France are first available at age 60. The age-specific rate of departure from the labor force in France jumps to approximately 60 percent at that age, as

shown in Figure 5. (The rather large departure rates before the early-retirement age reflect the guaranteed income provisions for employees who become "unemployed," even if they are not eligible for social-security benefits.) The collective evidence for all countries combined shows that statutory social-security eligibility ages contribute importantly to early departure from the labor force. In addition, unemployment and disability programs serve as early-retirement programs in many countries. This evidence is discussed in some detail in Gruber and Wise (1998) and in the individual country papers.

IV. The Implicit Tax on Work and Labor-Force Participation

The high rate of departure at the early-retirement age in France also illustrates the role of the implicit tax rate on work imposed by social-security plan provisions. At the early-retirement age in France, the implicit tax rate is nearly 70 percent for persons with median lifetime earnings. Such high tax rates are common in European countries, with tax rates over 50 percent in many instances, and in one case as high as 141 percent.

Drawing on the evidence from all of the country studies, we find that the relationship between the implicit social-security tax on work is strongly related to the labor-force participation of older persons. There is no completely satisfactory way to summarize the country-specific incentives for early retirement. The measure we use is based on continued labor earnings once a person approaches eligibility for social-security benefits. For this paper, we sum the implied tax rates on continued work beginning at age 55 and running through age 69. We call this the "tax force to retire." The measure ranges from less than 1 in Japan to over 9 in Italy. (A measure of 15 would imply a 100-percent tax rate on all earnings beginning at age 55.)

The relationship between this tax force to retire and unused labor-force capacity is shown in Figure 6, which presents a scatter plot of the tax force to retire and unused labor capacity between ages 55 and 65. The relationship is clear: there is a strong correspondence between the tax force to retire and

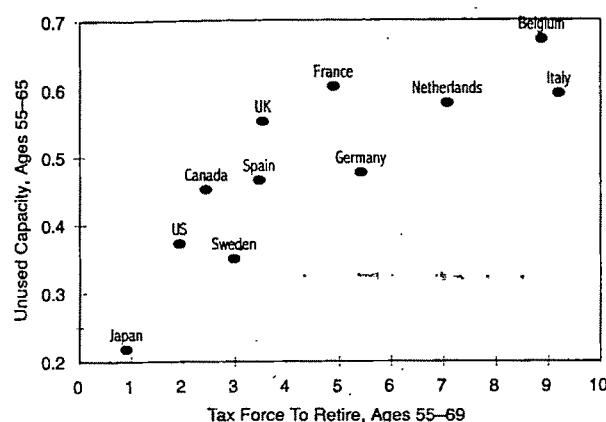


FIGURE 6. UNUSED LABOR-FORCE CAPACITY
VERSUS TAX FORCE TO RETIRE

unused labor capacity. The relationship is non-linear, however. If unused capacity is plotted against the logarithm of the tax force, creating an approximate linear relationship, the "fit" is surprisingly tight. A regression of unused capacity on the logarithm of the tax force, indicates that 82 percent of the variation in unused capacity can be explained by the social-security tax force to retire. Thus, these data suggest a strong relationship between social-security incentives to quit work and the labor-force departure of older workers. As shown in Gruber and Wise (1998), the relationship is not very sensitive to alternative age ranges for measuring either the tax force or unused capacity.

The correspondence between the two should be understood in a broader context, however. There are two distinct issues: First, while it seems apparent that social-security provisions do affect labor-force participation, it also seems apparent from the country papers that in at least some instances the provisions were adopted to encourage older workers to leave the labor force. For example, anecdotal evidence suggests that in some countries it was thought that withdrawal of older employees from the workforce would provide more job opportunities for young workers. This possibility does not by itself bring into question a causal interpretation of the relationship between plan provisions and retirement. To the extent that it is true, it simply says that in some instances the provisions were adopted for a

particular reason. And, the data show that they worked.

The second issue, however, must temper a causal interpretation of the results. It could be argued that, to some extent at least, the social-security provisions were adopted to accommodate existing labor-force participation patterns, rather than the patterns being determined by the provisions. For example, early-retirement benefits could be provided to support persons who are unable to find work and thus are already out of the labor force. While this is surely possible, the weight of the evidence suggests otherwise. The paper by Börsch-Supan and Reinhold Schnabel (1998) in this session and Blanchet and Pelé (1998) provide clear illustrations that changes in social-security provisions precede changes in labor-force participation, as do other examples in the individual country papers.

In short, it is clear that there is a strong correspondence between the age at which benefits are available and departure from the labor force. Social-security programs often provide generous retirement benefits at young ages. In addition, plan provisions often imply large financial penalties on labor earnings beyond the social-security early-retirement age. Furthermore, in many countries, disability and unemployment programs effectively provide early-retirement benefits before the official social-security early-retirement age. We conclude that social-security program provisions have indeed contributed to the decline in the labor-force participation of older persons, reducing the potential productive capacity of the labor force. It seems evident that if the trend to early retirement is to be reversed, as will almost surely be dictated by demographic trends, changing the provisions of social-security programs that induce early retirement will play a key role.

REFERENCES

- Blanchet, Didier and Pelé, Louis-Paul. "Social Security and Retirement in France." National Bureau of Economic Research (Cambridge, MA) Working Paper No. 6214, 1997; in Jonathan Gruber and David A. Wise, eds., *Social security programs and retirement around the world*. Chicago:

- University of Chicago Press, 1998 (forthcoming).
- Blundell, Richard and Johnson, Paul.** "Pensions and Retirement in the United Kingdom." National Bureau of Economic Research (Cambridge, MA) Working Paper No. 6154, 1997; in Jonathan Gruber and David A. Wise, eds., *Social security programs and retirement around the world*. Chicago: University of Chicago Press, 1998 (forthcoming).
- Boldrin, Michele; Jimenez-Martin, Sergi and Peracchi, Franco.** "Social Security and Retirement in Spain." National Bureau of Economic Research (Cambridge, MA) Working Paper No. 6136, 1997; in Jonathan Gruber and David A. Wise, eds., *Social security programs and retirement around the world*. Chicago: University of Chicago Press, 1998 (forthcoming).
- Börsch-Supan, Axel and Schnabel, Reinhold.** "Social Security and Declining Labor-Force Participation in Germany." *American Economic Review*, May 1998a (*Papers and Proceedings*), 88(2), pp. 173-78.
- _____. "Social Security and Retirement in Germany." National Bureau of Economic Research (Cambridge, MA) Working Paper No. 6153, 1997; in Jonathan Gruber and David A. Wise, eds., *Social security programs and retirement around the world*. Chicago: University of Chicago Press, 1998b (forthcoming).
- Brugiavini, Agar.** "Social Security and Retirement in Italy." National Bureau of Economic Research (Cambridge, MA) Working Paper No. 6155, 1997; in Jonathan Gruber and David A. Wise, eds., *Social security programs and retirement around the world*. Chicago: University of Chicago Press, 1998 (forthcoming).
- Diamond, Peter and Gruber, Jonathan.** "Social Security and Retirement in the United States." National Bureau of Economic Research (Cambridge, MA) Working Paper No. 6097, 1997; in Jonathan Gruber and David A. Wise, eds., *Social security programs and retirement around the world*. Chicago: University of Chicago Press, 1998 (forthcoming).
- Gruber, Jonathan.** "Social Security and Retirement in Canada." National Bureau of Economic Research (Cambridge, MA) Working Paper No. 6308, 1997; in Jonathan Gruber and David A. Wise, eds., *Social security programs and retirement around the world*. Chicago: University of Chicago Press, 1998 (forthcoming).
- Gruber, Jonathan and Wise, David A.** "Social Security Programs and Retirement Around the World: Introduction and Summary." National Bureau of Economic Research (Cambridge, MA) Working Paper No. 6134, 1997; in Jonathan Gruber and David A. Wise, eds., *Social security programs and retirement around the world*. Chicago: University of Chicago Press, 1998 (forthcoming).
- Kapteyn, Arie and de Vos, Klaas.** "Social Security and Retirement in the Netherlands." National Bureau of Economic Research (Cambridge, MA) Working Paper No. 6135, 1997; in Jonathan Gruber and David A. Wise, eds., *Social security programs and retirement around the world*. Chicago: University of Chicago Press, 1998 (forthcoming).
- Oshio, Takashi and Yashiro, Naohiro.** "Social Security and Retirement in Japan." National Bureau of Economic Research (Cambridge, MA) Working Paper No. 6156, 1997; in Jonathan Gruber and David A. Wise, eds., *Social security programs and retirement around the world*. Chicago: University of Chicago Press, 1998 (forthcoming).
- Palme, Mårten and Svensson, Ingemar.** "Social Security, Occupational Pensions, and Retirement in Sweden." National Bureau of Economic Research (Cambridge, MA) Working Paper No. 6137, 1997; in Jonathan Gruber and David A. Wise, eds., *Social security programs and retirement around the world*. Chicago: University of Chicago Press, 1998 (forthcoming).
- Pestieau, Pierre and Stijns, Jean-Philippe.** "Social Security and Retirement in Belgium." National Bureau of Economic Research (Cambridge, MA) Working Paper No. 6169, 1997; in Jonathan Gruber and David A. Wise, eds., *Social security programs and retirement around the world*. Chicago: University of Chicago Press, 1998 (forthcoming).

Social Security and Labor-Force Participation in the Netherlands

By ARIE KAPTEYN AND KLAAS DE VOS*

As in most developed countries, the population share of the elderly in the Netherlands is increasing. The share of the population over 65 grew from 8 percent in 1950 to 13 percent in 1995 and is expected to be 21 percent by 2050. If nothing else changes, this will cause a considerable increase in expenditures on social security (SS), which covers essentially all persons aged 65 or over.

A more immediate concern is the low labor-market participation rate of persons below age 65 and the costs of providing income to them. Well before the age of 65, individuals leaving the labor force may be eligible for various kinds of earnings-replacing benefits, such as disability insurance (DI), unemployment insurance (UI), welfare, and early retirement (ER).

Below we discuss the labor-market participation of the elderly and the incentives provided by these transfer programs.

I. Key Features of the Social Security System

SS was introduced in 1957. It is financed as pay-as-you-go. Each individual of age 65 or over (whether retired or not) is entitled to a flat-rate benefit of 50 percent of the statutory minimum wage, with a supplement of 20 percent for single persons, of 40 percent for single parents with a dependent child, and of up to 50 percent for persons with a partner younger than 65. Benefits are essentially independent of one's earnings history and are mostly not means-tested (except when one's partner is younger than 65).

Approximately 80 percent of households with a head over age 65 receive an occupational pension on top of SS. Typically, occu-

pational pensions and SS add up to 70 percent of final pay if one has worked for 40 years. After-tax replacement rates can go as high as 90 percent. Generally, if an employer offers a pension plan, participation is compulsory.

The main transfer programs for individuals below age 65 are welfare, DI, UI, and ER. Households with a head younger than age 65 and without other sources of income (and limited household wealth) are entitled to welfare. The level of benefits for a couple approximately equals the after-tax minimum wage.

DI (introduced in 1967, and expanded to include self-employed individuals in 1976) covers all employees and self-employed persons against loss of earnings due to long-term sickness and disability. Currently, DI benefits for employees who lost more than 80 percent of their earnings capacity start at 70 percent of previous earnings (up to a maximum). In contrast to welfare, DI benefits are not wealth-tested.

As there is extensive employment protection in the Netherlands, while at the same time DI arrangements are more generous than UI, both employers and employees have a preference for DI over UI. The rise in costs caused by the popularity of DI led the government to tighten entry conditions and to reduce benefit levels.

UI benefits mainly differ from DI in that they run out after a while or fall to the level of a welfare benefit. However, most people aged 60 or over who become unemployed will receive UI benefits equal to 70 percent of their previous earnings up to age 65. Furthermore, above age 57.5, unemployed individuals do not have to seek employment actively in order to qualify for UI, and hence they can de facto retire.

ER became increasingly common during the 1980's. Currently, many firms are trying to reduce the costs, by reducing the benefits or increasing the ER age. Typically, ER benefits amount to 70–80 percent of previous earnings

* CentER, P.O. Box 90153, 5000 LE Tilburg, The Netherlands. While writing this paper Kapteyn was visiting the Industrial Relations Section of Princeton University.

up to age 65. Again, after-tax replacement rates are higher.

ER is organized via one's occupational pension fund, or by the employer. It is mostly financed as pay-as-you-go. Eligibility usually requires ten years of employment as well as complete withdrawal from the labor market.

II. Labor-Market Participation of Older Persons¹

Since 1960, the labor-force participation of older men has declined in all age groups. The decline is particularly dramatic for 60–64-year-olds, from about 80 percent in 1960 to only 20 percent in 1994. For men aged 65 or over, labor-force participation fell from about 20 percent in 1960 to about 3 percent in 1985. Since then, Statistics Netherlands has stopped recording their participation rate.

Women's participation in the age group 45–54 increased from less than 20 percent in 1960 to more than 40 percent in 1994. The participation rates in the older age groups have remained low.

The relatively generous SS and pension system can explain why not many people work after age 65. In the younger age groups, the relatively generous DI scheme was the first program to offer an attractive way to retire. In particular, in periods with rapidly increasing unemployment in the 1970's and 1980's, the disability route to retirement became a popular alternative to general layoffs. In 1968, 12 percent of males between ages 55 and 64 received a DI benefit. From 1975 to 1985, this percentage increased from 21 percent to 37 percent.

In addition, as a way to shed labor during the 1980's many firms started to offer generous ER programs. In 1981, about 2 percent of the males between ages 55 and 64 received an

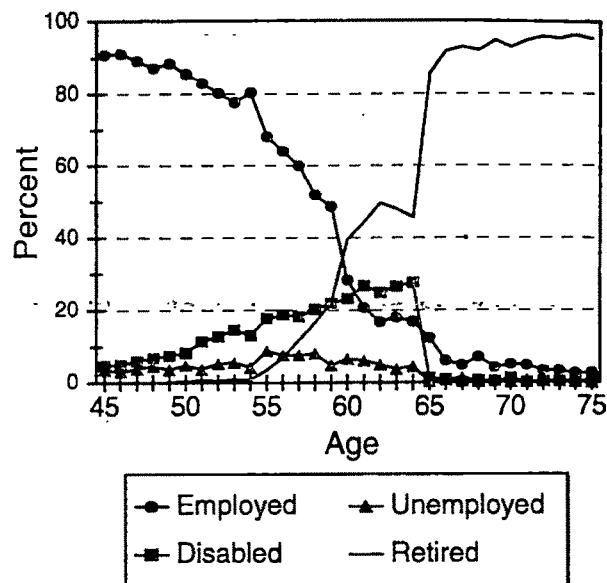


FIGURE 1. DISTRIBUTION OF ACTIVITIES OF MEN, BY AGE

ER benefit; in 1987, this percentage had increased to about 10 percent; and in 1995 to about 17 percent.

Figure 1 depicts male labor-market status across ages in 1994. One sees that, after age 50, male labor-force participation falls steeply. This fall is concomitant with a rise of the percentage of males on disability. After age 55, the percentage of males on disability keeps rising, and the percentage of (early) retirees rises from almost zero at age 54 to almost 50 percent beyond age 60. Beyond age 65, almost all men are retired. For females (not shown here) participation is only about half that of men. This is also true of the percentage of women on DI, which rises slightly from age 45 to age 64 but remains well below the corresponding percentage for men.

III. Retirement Incentives

For simplicity of terminology, we will denote by SSW ("social security wealth") the actuarially discounted sum of future benefits of all programs combined, minus the actuarially discounted sum of future contributions to these programs. To illustrate the retirement incentives of the current transfer programs, we will compute SSW and implicit tax/subsidy rates on working one additional year for some stylized cases. As in Jonathan Gruber and

¹ All data used to obtain the figures in this section come from Statistics Netherlands. Historical data are from the Census (1960, 1971) and Labor Force Surveys (1975, 1979, 1985, 1990, 1994). Figures are not adjusted for changes in definitions. For the situation in 1994, we use the 1993/1994 Housing Needs Survey (WBO), a nationally representative survey (55,000 households). The hazard rates given in Section III are based on pooling the 1984–1986 and 1992–1994 waves of the Socio-economic Panel (about 5,000 households) respectively.

David Wise (1997), tax/subsidy rates are defined as the change in the SSW relative to what an individual would earn over the coming year.

In the calculations the following assumptions are made: (i) benefits remain constant in real terms after 1995; (ii) mortality rates (taken from Statistics Netherlands [1992]) of a worker and his spouse are independent; (iii) the real discount rate is 3 percent; (iv) to compute net benefit and pension levels, payroll and income taxes are subtracted; (v) for the years after 1995, the tax schedule for 1995 applies; (vi) between 1985 and 1995, the wage moved with the statutory minimum wage. We consider a man born in January 1930, with median earnings in 1985. His wife is three years younger and has no earnings.

To appreciate the various factors influencing the taxes or subsidies on working, consider a worker between his 55th and 65th birthday who may or may not be working another year. Working another year may affect SSW in different ways. (i) When working one owes payroll taxes towards SS, UI, DI, and possibly, private pensions. This lowers SSW. (ii) The worker may forgo a year of benefits (DI, ER if aged above 60), which lowers SSW. (iii) At age 59, the worker would be entitled to ER if he worked another year. This would considerably increase SSW. (iv) Working another year implies accumulating another year of occupational pension rights, which increases SSW.

Table 1 provides results of our calculations for two cases. The first one (Table 1A) concerns a person who is entitled to an ER pension when retiring at age 60 or later. If he leaves the labor force before turning 60, he loses that right and also will stop accumulating pension rights. Typically, ER benefits are 80 percent of previous earnings; the after-tax replacement rate is about 90 percent. At age 65, the occupational pension supplements SS to 70 percent of final pay, which likewise results in an after-tax replacement rate of about 90 percent.

The high replacement rate, in combination with the high tax on a continuation of work, provides a powerful incentive to retire as soon as the worker is eligible for ER. It should be noted that in reality ER eligibility ages vary,

TABLE 1—INCENTIVE CALCULATIONS

Age at last year of work	Replacement rate	SSW ^a	Accrual ^a	Tax/subsidy
A. SS + ER + PP:				
54		266,958		
55		247,365	-19,593	0.687
56		229,033	-18,332	0.650
57		212,121	-16,912	0.612
58		196,668	-15,453	0.578
59	0.910	296,367	99,699	-3.777
60	0.906	258,463	-37,903	1.410
61	0.900	222,715	-35,748	1.384
62	0.902	188,559	-34,157	1.339
63	0.892	157,316	-31,242	1.280
64	0.909	128,554	-28,762	1.222
B. SS + DI + PP:				
54	0.791	459,325		
55	0.789	417,164	-42,161	1.478
56	0.787	376,878	-40,285	1.428
57	0.788	338,751	-38,128	1.379
58	0.782	303,010	-35,741	1.338
59	0.761	269,520	-33,490	1.269
60	0.761	237,690	-31,830	1.184
61	0.759	207,718	-29,972	1.160
62	0.762	179,121	-28,598	1.121
63	0.758	152,290	-26,831	1.099
64	0.909	128,554	-23,735	1.009

^aUnits are Dutch guilders.

and hence the incentives to retire will vary across workers.

The second case (Table 1B) concerns someone who is entitled to DI if he stops working before age 65. The DI benefits amount to 70 percent of previous earnings, or almost 80 percent after tax. By working an additional year, the worker forgoes a year of DI benefits and pays an additional year of contributions. On the plus side, he accumulates additional pension rights. The resulting decrease in SSW amounts to an implicit tax rate on net earnings of more than 100 percent. So again, the incentives to exit the labor force appear to be strong.

In Kapteyn and De Vos (1997) we have also computed implicit tax rates for different types of workers and using different assumptions. As in the cases considered above, these calculations imply that the Dutch system of SS, DI, ER, and occupational pensions provides strong incentives to retire as soon as possible, or when the worker is entitled to ER, at the ER eligibility age.

One can further illustrate the importance of incentives by comparing escape routes in the mid-1980's, when ER was still on the rise, and the mid-1990's. Around 1985, the yearly hazard rates into ER and DI for male workers between ages 59 and 61 were, respectively, 0.19 and 0.04. Around 1992, the hazard rate into ER was 0.31, and into DI, 0.02. These numbers illustrate a substantial increase in the ER hazard over a relatively short period and suggest a slight fall in the corresponding hazards for DI.

IV. Concluding Remarks

The dramatic fall in labor-force participation among the elderly in the Netherlands can be explained largely by the introduction over the past four decades of a number of new arrangements which created incentives to retire.

These incentives exacerbate the economic consequences of the demographic shift, as fewer workers have to support an ever increasing number of nonparticipants in the labor market. In comparison with other countries, the Netherlands has one advantage in that the

occupational pensions are fully funded. Pension-fund reserves in the Netherlands are about 130 percent of GDP, by far the highest in the OECD. This will make the financing of retirement in the future less problematic than in most OECD countries.

REFERENCES

- Gruber, Jonathan and Wise, David. "Social Security Programs and Retirement Around the World." National Bureau of Economic Research (Cambridge, MA) Working Paper No. 6134, August 1997; in Jonathan Gruber and David Wise, eds., *Social security programs and retirement programs around the world*. Chicago: University of Chicago Press, 1998 (forthcoming).
- Kapteyn, Arie and de Vos, Klaas. "Social Security and Retirement in the Netherlands." National Bureau of Economic Research (Cambridge, MA) Working Paper No. 6135, August 1997.
- Statistics Netherlands. *Life tables for the Netherlands, 1986-1990*. Voorburg/Heerlen, Netherlands: Statistics Netherlands, 1992.

Pensions and Labor-Market Participation in the United Kingdom

By RICHARD BLUNDELL AND PAUL JOHNSON*

Unlike most other European countries the United Kingdom's pension system is not well described by an analysis of the social-security element. For 30 years or more, around half the workforce has been covered by private occupational pensions. Something like half of the income of pensioners comes from non-social-security sources, and this proportion is growing. Of the workforce in the mid-1990's, 75 percent are "contracted out" of the second-tier State Earnings Related Pension Scheme (SERPS) into private occupational or personal pensions.¹ Partly as a result of these facts the United Kingdom also differs from many other countries in one other important respect: its state pension system is solvent. Tax rates necessary to pay for it are not predicted to rise despite the fact that the number of people over state retirement age is predicted to rise from 15.7 percent of the whole population to over 24 percent in 2050.²

In this paper we begin by describing the labor-market behavior of individuals around pension age. We also consider the coverage of the various parts of the social-security system. We go on to explain the structure of state pensions in the United Kingdom and compute the incentives for retirement that the structure creates, focusing on early retirement and the role of disability benefits, which are especially relevant for lower-skilled workers. We compare

these incentives with those facing workers with private occupational schemes who now make up the majority of older workers close to retirement. The structure of incentives is found to match well with the observed patterns of labor-market participation in the data.

I. The Labor-Market Behavior of Older Persons in the United Kingdom

The labor-market behavior of older persons in the United Kingdom has been characterized by a severe fall in the participation of men. The rate of participation among recent cohorts falls sharply below 80 percent after the age of 50 and declines rapidly thereafter. In contrast the secular rise in the participation of women has resulted in a small upward trend in participation among women in the 55–60 age bracket, with participation rates approaching those for men in that age group. The vast majority (80 percent or so) of men in their late forties are (full-time) workers. This proportion falls steadily, reaching 75 percent of those in their early fifties, dropping to 60 percent of those in their late fifties, with a sharp drop to 40 percent of 60-year-olds. This drops again to 30 percent by age 64 and then under 10 percent at 66. Hazard rates presented in Blundell and Johnson (1997) show a growing rate of exit for men beginning in their early fifties. For women the pattern is similar but with much lower full-time working and higher levels of part-time work. Work participation among women tails off quite rapidly for the 50-year-old women, falling from about 60 percent in the late forties to 30 percent in the late fifties and 20 percent at age 60. Given that the state pension becomes available at age 60 for women the increase in inactivity at that age is not surprising.

II. Key Features of the U.K. Pension and Social-Security System

It is hard to consider the U.K. state pension system in isolation from the private sector.

*Blundell: Institute for Fiscal Studies and Department of Economics, University College London, Gower Street, London, WC1E 6BT United Kingdom; Johnson: Institute for Fiscal Studies, 7 Ridgmount Street, London, WC1E 7AE United Kingdom. We thank Jonathon Gruber, Sarah Tanner, Jayne Taylor, and David Wise for their advice and comments. Financial support from the U.K. Economic and Social Research Council and the Centre for the Microeconomic Analysis of Fiscal Policy at the Institute for Fiscal Studies is gratefully acknowledged.

¹ See Andrew Dilnot et al. (1994) for an overview of the U.K. pension system.

² These figures include the effect of the equalization of state pension age at 65 for men and women, a change that will be phased in between 2010 and 2020.

Among recently retired pensioners, private pensions make up almost half of total income in retirement, with the mean occupational pension payment (among those receiving some payment) approaching £90 (\$144) per week for a single person in 1996, which compares with a flat-rate basic state pension of £61.15 per week. This basic pension level represents just 16 percent of average male earnings. With indexation in line with the Retail Price Index its level relative to earnings is falling; it was 20 percent of the male average in the late 1970's.³

Although unrelated to earnings levels, the basic pension is nonetheless a "contributory" benefit, at least in principle. Entitlement to full benefit depends on contributions being made for 90 percent of a working life. These contributory conditions are not at all so onerous as they appear. Any time spent unemployed or sick/disabled attracts credits, which count in just the same way as contributions, and since 1978, time spent looking after children has reduced the effective number of years of contributions required through a system called Home Responsibilities Protection (HRP). Virtually all men aged 65 and over receive a full basic pension on the basis of their own contributions. The coverage of women currently over 60 is less comprehensive. However, these low rates of entitlement among married women reflect long periods spent out of the labor market by older cohorts. As a consequence of the increasing labor-market participation at younger ages and the home-responsibility rule, by the early years of the next century the vast majority of women will retire with entitlement to a full basic pension.

It is possible to defer pension receipt by up to five years. Deferral results in an increase in pension entitlement of 7.5 percent per year. This is more valuable to women than to men because of their higher life expectancy. Possibly as a result of this, 17 percent of female pensioners and 11 percent of males receive increments to their basic pensions as a result of deferral. Deferral is becoming less widespread

following the abolition, in 1989, of the "earnings rule," which effectively meant that those (women aged 60–64, and men 65–69) earning more than £75 per week (in 1989) had their pension entitlement severely reduced. The reduction was 50p for every £1 between £75 and £79 of earnings and £1 for every £1 thereafter. Virtually all those affected deferred their pension receipt rather than taking a reduced amount. Its importance is reflected in the fact that nearly a quarter of men and a third of women over age 80, with pension entitlement in their own right, have pension increments as a result of deferral. One might have expected the complete abolition of this rule to lead to significantly changed behavior among those in the relevant age ranges. In Blundell and Johnson (1997) we present new results based on earnings distributions in 1987–1988 and in 1991–1992 that support this hypothesis. There is clear evidence of bunching at the earnings-rule level in 1987–1988. There is no such peak in the later data.

Within the state welfare system itself, the most dramatic changes with respect to numbers receiving benefits have been in the number of pre-pension-age individuals receiving benefits initially designed for the long-term sick and disabled. The Incapacity Benefit (previously the Invalidity Benefit [IVB]) is the most important of these. It is a contributory benefit payable to long-term sick individuals who can show they are incapable of work due to illness or disablement and have been so for at least 28 weeks. Given that about a quarter of all men aged 60–64 (and nearly 20 percent aged 55–59) were in receipt of IVB in 1994, there can be little doubt that IVB has been used as an early-retirement vehicle.

The contributory basic benefit system was originally designed as a purely flat-rate arrangement intended only to provide a bare minimum income level. SERPS was introduced in 1978, with the intention that it would start paying out full benefits 20 years hence. The result of the introduction of SERPS, especially for the generation retiring in the years around 2000, will be to increase significantly the social-security income (and thus total income) of those without a private pension.

The retirement income of those in occupational pension schemes has been largely

³James Banks et al. (1996) use these changes in replacement rates across cohorts to assess the consumption-smoothing hypothesis at retirement.

unaffected by the introduction of SERPS. Individuals in schemes that guarantee a certain level of benefit can give up rights to SERPS and pay lower National Insurance contributions as a result. Since 1988, not only have traditional final-salary occupational pensions been able to contract out, but so also have group money-purchase and personal-pension schemes. So now about three-quarters of eligible workers (i.e., those earning more than a lower earnings limit set for SERPS contributions) are not covered directly by SERPS. Half are in occupational schemes, and another quarter are in personal pensions.⁴

III. Retirement Incentives in the State System

In this section we consider the retirement incentives within the U.K. social-security system. The analysis reveals a number of interesting features of the U.K. retirement-benefit system. In particular it demonstrates the role of benefits available before state pension age. In our simulations we consider the incentives facing a married man born in 1930 and so reaching state retirement age of 65 in 1995. We consider two scenarios: the first ignores the impact of the invalidity benefit on early-retirement incentives, while the second incorporates this into the calculations. In each simulation we compute the amount of earnings-related and basic pension to which the individual would be entitled in the first and subsequent years of retirement. Blundell and Johnson (1997) provides specific details and simulations for a number of other scenarios. Given life tables and an assumed discount rate (3 percent), we calculate an expected net present discounted value of social-security wealth, SSW.

For the first scenario, in each year up to age 65 the accrual of SSW is slightly negative. Net SSW conditional upon retiring at age 65 is about £2,000 less than that conditional on retiring at age 55. This difference is small. It reflects two features of the U.K. system. Until age 65 the individual will be paying 10 percent

of his earnings in contributions each year, and his employer will be paying an additional 10 percent. So the cost to working an extra year is substantial (we use observed median earnings). The benefit of working an extra year, in terms of SSW, comes through extra earnings-related SERPS being accrued. Basic pension entitlement, which makes up the greater part of the total state pension, is unaffected by extra years of work. The loss in net income to higher contributions is significant for each extra year of work but only adds on once. The extra amount of SERPS earned is small for each year but payable for many years, especially given the existence of a younger wife. These values come close to canceling each other out, but the negative effect of extra contributions is just the greater.

The impact of introducing the disability benefit is dramatic. Assume that the benefit is available at age 60; then each extra year of work means forgoing a full year's benefits with only a small future increase in SERPS as compensation. The effects of an extra year of work are to reduce SSW by about £8,000 per year. This is equivalent to a tax rate of more than 70 percent on the year's earnings and means a fall in SSW of around 10 percent or more for each year of work. The penalty for staying on in work can be great indeed. Moreover, to the extent that individuals are able to claim invalidity benefits before age 60 (or receive income support without a work test) these arguments extend backwards even further. Of 55–59-year-old men, nearly 20 percent in 1994 were receiving the Invalidity Benefit, a number that has doubled since the early 1980's. It is also the case that until the beginning of the 1990's SERPS additions were payable in respect of invalidity pensions as well as in respect of retirement pensions. For low earners the effects of being eligible for benefits are even more spectacular. (Those unable to qualify for IVB may be eligible for means-tested benefits at a similar level.) In sum, the potential incentives for older low-to-middle earners without private pensions to leave the labor market are very considerable.

These observations raise interesting issues about the structure of the U.K. benefit system and appear to fit rather well with the observed behavior of many older men. Put together with

⁴ Very few personal pensions have reached maturity. In the future, personal pensions will become much more important in this context (see Dilnot et al., 1994).

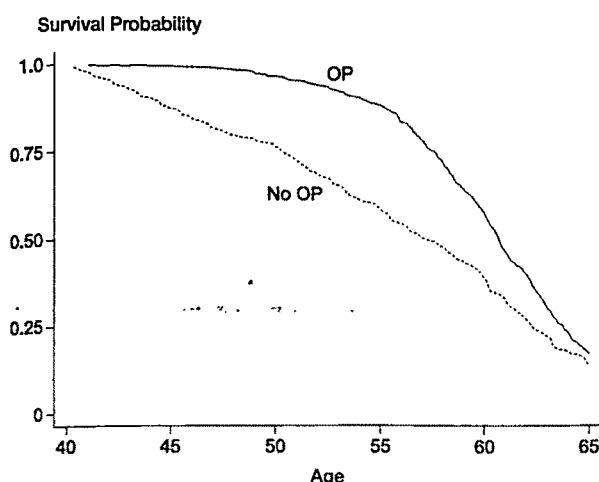


FIGURE 1. SURVIVAL FUNCTIONS FOR MEN
(PROBABILITY OF REMAINING IN EMPLOYMENT)

Note: OP = occupational pension.

the fall in demand for lower-skilled workers, the relative generosity of social security for older groups, especially through apparently easy access to the invalidity benefit, could well help explain much of the observed fall in participation rates among older less-skilled workers. Indeed there is some evidence for relatively elastic labor-supply behavior among older men in the United Kingdom (see the results using the U.K. Retirement Survey data in Costas Meghir and Edward Whitehouse [1997] and Richard Disney et al. [1994]).

IV. Retirement Incentives and Occupational Pensions

As we have stressed throughout, for a large part of the population, social-security pensions play only a secondary role in providing retirement income. In the private sector the standard occupational pension offers a pension equal to one-60th of final salary for each year of membership in the scheme. When early retirement is available it is often available on generous terms that clearly result in losing pension wealth by working longer. The only group for whom this is unlikely to be true are those who might expect substantial pay increases in the years approaching normal retirement age.

The differing nature of the rules governing occupational pension schemes and those governing state pensions clearly induce different

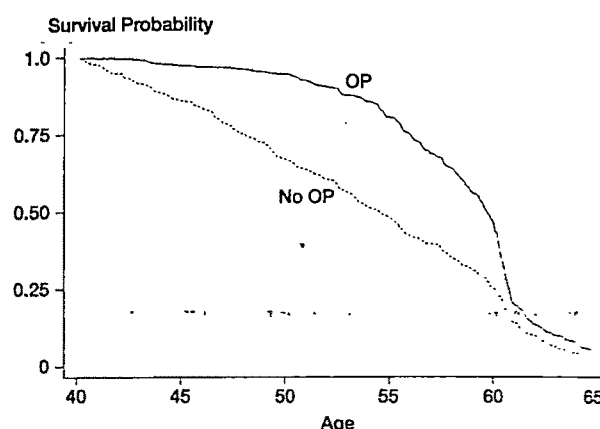


FIGURE 2. SURVIVAL FUNCTIONS FOR WOMEN
(PROBABILITY OF REMAINING IN EMPLOYMENT)

Note: OP = occupational pension.

incentives to retire before the standard retirement age. On becoming eligible, which typically occurs after age 55, the most obvious impact of occupational pension schemes operates through a wealth effect. Individuals eligible for early retirement are less likely to work when their pension income is higher. However, occupational pensions also may give an incentive to work longer since continued employment increases eventual pension entitlement, when pensions are typically linked to final earnings.

These differential incentives should show themselves in observed transition rates out of employment for those nearing retirement. To analyze this, Blundell and Johnson (1997) considers results from the U.K. Retirement Survey. This data source covers some 2,500 households in the age range 55–69.⁵ Survival functions (i.e. the age-specific probability of remaining in employment) for men and women are presented in Figures 1 and 2. The labor-market experience of the two groups is shown to be profoundly different. These survival functions confirm the importance of the incentives provided by occupational pension schemes: the survival

⁵ It gives detailed employment and pension life histories. It is a retrospective work-history data set unique in the United Kingdom, which records all job spells for each individual in the household.

probability is considerably higher just before retirement benefits may become due (either "full" or early retirement), and thereafter the survival probability falls much more rapidly than that of those not covered by pension schemes. The differences in these survival-to-retirement functions are consistent with what we would expect given the different incentive structures.⁶ Those without an occupational pension tend to be less skilled and have not seen any increase in real wages. The incentives for retirement before age 55 for this group are greater than for those with an occupational pension, as is portrayed in Figures 1 and 2 (see Meghir and Whitehouse [1997] for detailed elasticity estimates). Those with occupational pensions start to be able to take attractive levels of pensions from age 55, and this is also clearly reflected in the survival rates.

V. Conclusions

There have been significant changes in labor-market behavior among older individuals in the United Kingdom since the 1970's. Participation and activity rates, especially among men over age 55 have fallen dramatically. For those who can get invalidity benefits, who account for more than 40 percent of nonworking 60–64-year-olds, the system comes close to working as providing early-retirement benefits with no actuarial reduction. The relative generosity of these benefits and the incentives which they create, combined with the reduced demand for unskilled labor, play an important part in explaining the observed fall in labor-market participa-

tion. Among those with occupational pensions, significant increases in pension wealth have had an important effect on increased early retirement, as has the relatively generous treatment of early retirement by many occupational schemes. The retirement behavior of those individuals with occupational schemes has been shown to differ significantly from that of those without occupational pensions in ways that are consistent with the incentives underlying the different schemes.

REFERENCES

- Banks, James; Blundell, Richard and Tanner, Sarah. "Is There a Savings and Retirement Puzzle?" University College London Discussion Paper in Economics 96-25, 1996; *American Economic Review* (forthcoming).
- Blundell, Richard and Johnson, Paul. "Pensions and Retirement in the UK." National Bureau of Economic Research (Cambridge, MA) Working Paper No. 6154, September 1997; in Jonathan Gruber and David Wise, eds., *Social security programs and retirement around the world*. Chicago: University of Chicago Press, 1998 (forthcoming).
- Dilnot, Andrew; Disney, Richard; Johnson, Paul and Whitehouse, Edward. *Pensions policy in the UK: An economic analysis*. London: Institute for Fiscal Studies, 1994.
- Disney, Richard; Meghir, Costas and Whitehouse, Edward. "Retirement Behavior in Britain." *Fiscal Studies*, August 1994, 15(1), pp. 24–43.
- Meghir, Costas and Whitehouse, Edward. "Labor Market Transitions and Retirement of Men in the UK." *Journal of Econometrics*, September 1997, 79(2); pp. 327–54.

⁶ The differences will also, in part, reflect endogenous sorting by individuals.

Social Security and Declining Labor-Force Participation in Germany

By AXEL BÖRSCH-SUPAN AND REINHOLD SCHNABEL *

Germany has one of the most generous retirement systems in the world. At the same time, Germany also faces one of the most incisive population aging processes. The ratio of workers to pensioners will decrease to less than 1:1 within the next generation. This will put the German pay-as-you-go social-security system under severe pressure. Already, Germany has experienced a sharp increase in the contribution rate to the social-security system.

This paper shows that the German public pension system is a textbook example of negative incentive effects on system participation and old-age labor supply. The design of the current system will make coping with the future demographic challenges particularly difficult.

Incentive effects of private U.S. pension plans on retirement behavior have been surveyed by Lawrence Kotlikoff and David Wise (1987) and formalized in the option-value analysis by James Stock and David Wise (1990). Costas Meghir and John Edwards (1993) exploit the "opt out" mechanism of the state earnings-related pension system in the United Kingdom. What makes the German case particularly interesting is the universality of the German public pension system: its incentive effects influence almost all workers' behavior.

I. The German Public-Pension System

Germany has the oldest formal social-security system, introduced in 1889 by Chancellor Otto von Bismarck. Originally a fully funded disability insurance, it soon became a mandatory retirement insurance system and was converted to a pay-as-you-go

(PAYG) scheme after its capital stock had been severely eroded during World War II. In the course of the 1960's and 1970's, the German system evolved to a universal and very generous pension program both in terms of its replacement rate and its early-retirement provisions.

The German public pension system¹ is almost universal for two reasons. First, it is mandatory for every worker except for the self-employed and those with very small labor incomes. Because almost all German workers have been dependently employed at least at some point in their working career, almost every worker has a claim on a public pension. Second, the system has a very high replacement rate, generating net retirement incomes that are currently about 70 percent of preretirement net earnings for a worker with a 45-year earnings history and average lifetime earnings. This is substantially higher than the corresponding U.S. net replacement rate of about 53 percent. In addition, it provides relatively generous survivor benefits that constitute a substantial proportion of the total pension liability. Social-security income represents about 80 percent of household income of households headed by a person aged 65 and over, the remainder about equally divided among firm pensions, asset income, and private transfers. On the aggregate level, public pensions are 10.6 percent of GDP, a share more than 2.5 times larger than in the United States.

The German public pension system provides *old-age pensions* for workers aged 60 and older, *disability benefits* for workers below age 60, which are converted to old-age

* Börsch-Supan: Department of Economics, University of Mannheim, D-68131 Mannheim, Germany, CEPR, and NBER; Schnabel: Department of Economics, University of Mannheim, D-68131 Mannheim, Germany.

¹ The institutional description refers to the "Gesetzliche Rentenversicherung." Some branches have their own but similar PAYG retirement systems (e.g., civil servants, miners). For a detailed description see Börsch-Supan and Schnabel (1998).

pensions by age 65, and *survivor benefits* for spouses and children. In addition, preretirement (i.e., retirement before age 60) is possible using other parts of the public transfer system, mainly unemployment compensation. A main feature of the German old-age pensions is "flexible retirement" from age 63 for workers with a long service history. In addition, retirement at age 60 is possible for women, the unemployed, and workers who cannot be appropriately employed for health or labor-market reasons.

Benefits are computed on a lifetime contribution basis and adjusted according to the type of pension and retirement age. They are the product of four elements: (i) the employee's relative wage position, averaged over the entire earnings history, (ii) the number of years of service life, (iii) adjustment factors for pension type and (since the 1992 reform) retirement age, and (iv) the average pension level. The first three factors make up the "personal pension base" which is calculated when entering retirement. The fourth factor determines the income distribution between workers and pensioners in general and is adjusted annually to net wages. Thus, productivity gains are transferred each year to all pensioners. Due to a generous exemption, social-security benefits are tax-free unless income from other sources is high.

Roughly 80 percent of the budget of the German public pension system is financed by contributions, the rest by federal government revenue. Contributions are collected like a payroll tax, levied equally on employees and employers. The tax rate in 1998 is 20.3 percent of monthly gross income, and the tax base is capped at about 180 percent of average wages.

II. Disincentives on System Participation

The pay-as-you-go system started in 1957 and has produced fairly large rates of return during maturation. Labor-productivity increases were exceptionally high during the German "economic miracle," and the labor force has been steadily increasing. However, these golden times are gone. Labor productivity now increases at a rate only slightly higher than in the United States, and the size of the labor force has been falling since 1992.

Demography is the main threat to the German public pension system. The share of German elderly will increase from 21 percent in 1995 to 36 percent in the year 2035, when the aging process will peak. This will be the highest share in the world (a distinction shared with Switzerland and Austria). The old-age dependency ratio will far more than double from 21.7 percent in 1990 to 49.2 percent in 2030.

The tax base for the German PAYG system will therefore severely erode, and there is nothing that can rescue the generosity of the current PAYG system. Holding the current replacement rate and labor-participation patterns constant, the demographic change implies an increase of social-security contributions from 21 percent in 1998 to 34 percent in 2035. Adding this to an already high tax burden and increasing contributions to mandatory health and long-term care insurance, a policy of tax increases appears unsustainable. While an increase in female labor-force participation and higher immigration may help in the short run, the magnitude of the demographic change requires a substantial shift in retirement age combined with a severe cut in the replacement rate.

The main insight is that whatever the future policy mix may be (consisting of a tax increase, a benefit reduction, and a shift in retirement age), the rate of return will invariably go down. Figure 1 shows projections by Schnabel (1997) based on the official population projection and labor-force assumptions detailed in Börsch-Supan (1998) that bracket the real rate of return by birth cohort between the two extreme policies which either put the entire burden on the younger generation or on the elder generation. The real rate of return will be below 1 percent from about birth cohort 1950 on, and negative for cohorts born after 1980. This creates a huge disincentive problem for participation.

There is mounting evidence that these disincentive effects induce German workers to vote with their feet against the PAYG system by increasingly using exceptions from mandatory participation. The share of self-employed has gone up since 1991, and the number of part-time jobs with salaries below the social-security contribution threshold have

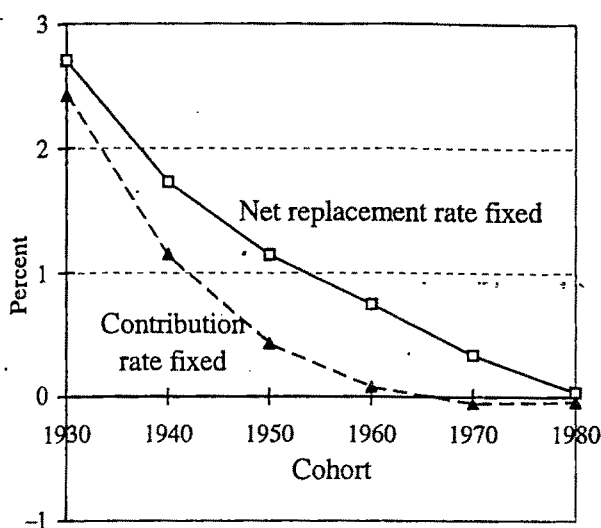


FIGURE 1. REAL RATE OF RETURN ON THE GERMAN PAYG SYSTEM

Source: Schnabel (1997).

increased dramatically at the same time. The clearest evidence of voting with the feet occurs among the self-employed who can choose between participation in the public pension system and market-based pension plans. Figure 2 shows that the proportion of male self-employed who actively contributed to the public system decreased from 62 percent in 1985 to 22 percent in 1995.² This reduced participation exhibits a strong age and cohort effect. Self-employed individuals who are age 45 and younger added less than a quarter of what self-employed individuals aged over 55 have added to the minimum contribution.

III. Incentives for Early Retirement

Before 1992, there was no adjustment of benefits when a worker retired earlier than at age 65. Only because benefits are proportional to the years of service, a worker with fewer years of service would get lower benefits. With a constant income profile and 40 years of service, each year of earlier retirement decreased pension benefits by 2.5 percent, much less than

² Active contribution refers to contributions that exceed the minimum amount (about \$50 per month at purchasing power parity) necessary to maintain a claim on the minimum pension and on disability pensions.

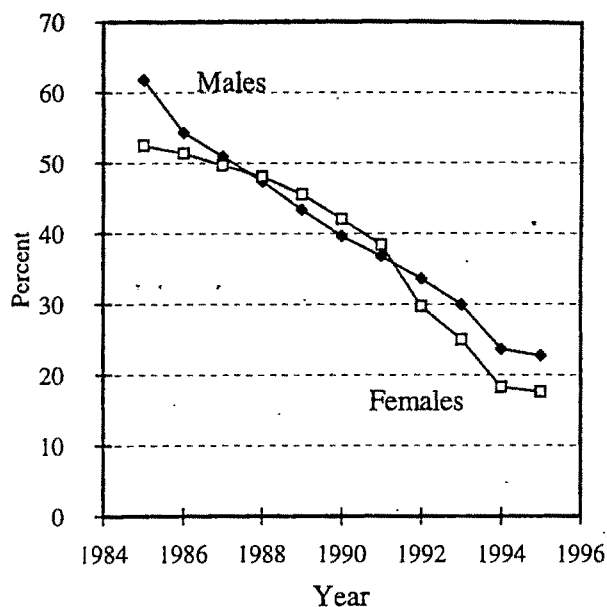


FIGURE 2. ACTIVE PARTICIPATION IN THE GERMAN PUBLIC PENSION PLAN AMONG SELF-EMPLOYED INDIVIDUALS

Source: Schnabel (1997).

the actuarial adjustment, which increases from about 5.5 percent at age 60 to 8 percent at age 65. The 1992 social-security reform is gradually changing this by introducing retirement age-specific adjustment factors to the benefit formula, but they remain about 2-percent below those required for actuarial fairness.

The lack of actuarial fairness creates a negative accrual of pension wealth during the early-retirement window at a rate reaching -9 percent when retirement is postponed from age 64 to age 65 (Fig. 3). Expressed as a percentage of annual labor income, this loss corresponds to a tax which exceeds 50 percent. After 2004, when the 1992 reform will have phased in, the negative accrual rate will reach -5 percent, corresponding to an implicit tax rate of almost 30 percent when retirement is postponed by one year at age 64.

The labor-supply disincentive for older workers created by this implicit tax is reflected in the data. Male labor-force participation plunges after age 55, when it is almost 90 percent, to 38 percent at age 60. It is less than 8 percent at age 65. Figure 4 shows the cross-sectional distribution of retirement ages that has its maximum at age 60, the earliest age at which retirement due to health and

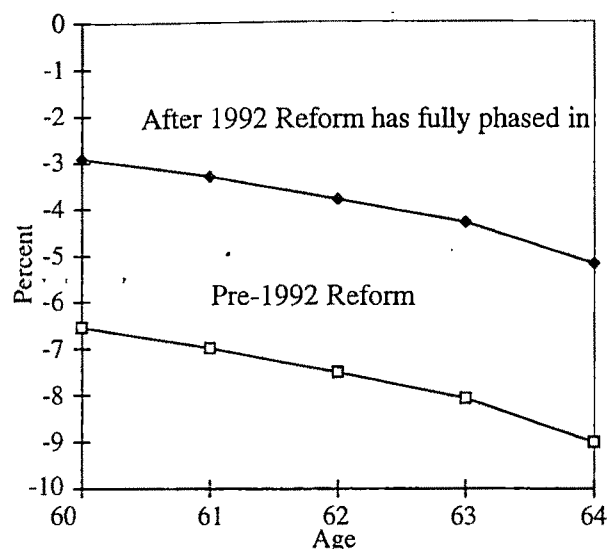


FIGURE 3. ACCRUAL RATES OF PENSION WEALTH BY RETIREMENT AGE

labor-market reasons without formal claims to disability benefits is possible. The other spikes correspond to age 63 for flexible retirement and to age 65 for workers with short work histories.

The incentive effects are even stronger if one manages to claim disability status for health or labor-market reasons, because no adjustments apply (not changed in the 1992 reform). Thus, implicit tax rates are similar to the pre-1992 regime, in excess of 60 percent for workers retiring before age 60. Disability is an important pathway to retirement. In 1981, at its peak, 68 percent of male workers retired through the pathway of disability benefits. Since then, disability eligibility was tightened, but still today more workers enter retirement through disability insurance (41 percent in 1995) than through regular old-age pensions (35 percent). In addition, "preretirement" schemes that combine severance pay and unemployment benefits with the early-retirement provisions for unemployed workers account for 24 percent of retirement entries.

Formal econometric analyses confirm that pension rules strongly affect retirement behavior. Börsch-Supan (1992), Sikandar Siddiqui (1995), and Börsch-Supan and Peter Schmidt (1996) exploit cross-sectional and time-series variation in the option value of postponing retirement to estimate the incentive effects of the

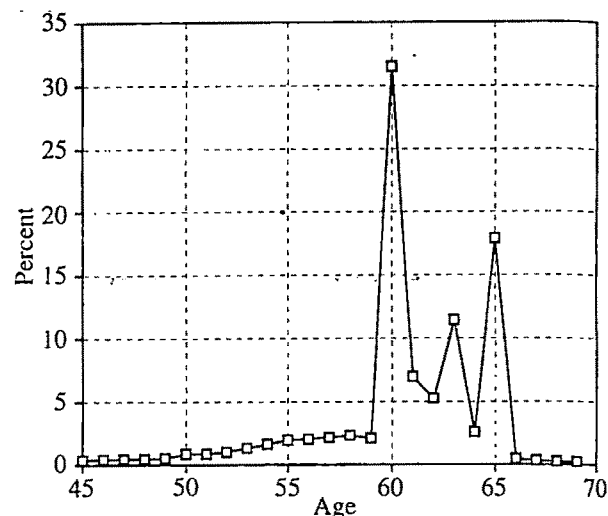


FIGURE 4. DISTRIBUTION OF RETIREMENT AGES (MEN)

German social-security system on early retirement. These studies use several variants of the option-value models compared in Robin Lumsdaine et al. (1992) and produce very robust results. The estimated coefficients can be used to compute the effect of the non-actuarial benefit adjustments on retirement age. According to Börsch-Supan and Schmidt (1996), the lack of actuarial fairness induces a shift of more than two years toward earlier retirement. The effects are most powerful for very early retirement (i.e., retirement before the official window period through disability or preretirement schemes). A shift to an actuarially fair system would cause retirement at ages 59 and below to drop from currently 32 percent to less than 18 percent. These potential effects are much larger than those simulated for the 1992 pension reform. Using the adjustment factors that will eventually be introduced by the 1992 reform, average retirement age is expected to increase by only about half a year, and very early retirement (before age 60) is expected to decline from 32 percent to only slightly above 28 percent.

The most convincing evidence that the early-retirement incentives are indeed causal for the low labor-force participation among older German workers is the "natural experiment" of the 1972 pension reform that introduced the "flexible retirement" option described in Section I. Before 1972, the earliest retirement age was 65 except for disabled

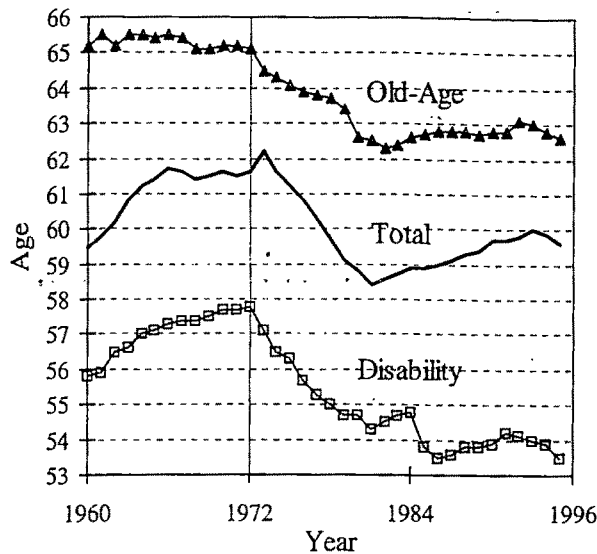


FIGURE 5. AVERAGE RETIREMENT AGE

workers. Figure 5 shows the sudden decrease in the average retirement age after the 1972 reform.³ Average retirement age dropped from about 61.5 in 1971 to 58.5 in 1981 and has remained below age 60 since then. The spike in the 1973 average retirement age is due to an interesting composition effect also induced by the reform. In 1973, when the "flexible retirement" option was opened, the share of retirees entering through disability retirement decreased sharply, while the share of retirees claiming old-age pensions increased. At the same time, average retirement age dropped in both the old-age and the disability branch of the public pension system.

The effect of the 1972 reform is also clearly visible in the quickly shifting distribution of retirement ages. Figure 6 shows that in 1970 age 65 was the retirement age, while in 1975, about half of the retirees preferred to retire earlier. Five years later, the pattern of today (cf. Fig. 4) started to emerge.

IV. Conclusions

The responsiveness of both active participation in the public pension system and retire-

³ Average retirement age in a given year is the average age of those workers receiving public-pension income for the first time.

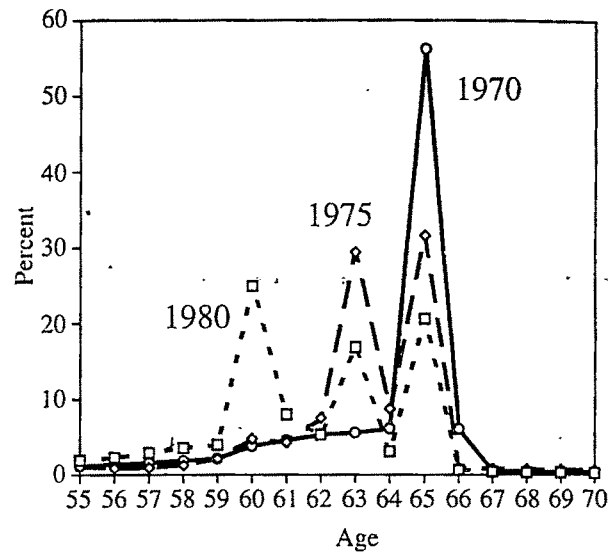


FIGURE 6. DISTRIBUTION OF RETIREMENT AGES, 1970, 1975, AND 1980

ment behavior to the incentives offered by the pension system has strong policy impacts. Rather than rewarding later retirement to reduce the negative effects on the German economy precipitated by quickly rising social-security taxes, social-security regulations in Germany have encouraged early retirement and thus aggravated the imbalance between the number of workers and pensioners in times of population aging. The 1992 German social-security reform has failed in this respect: it will only moderately decrease the distortions when fully phased in by the year 2004. Its effect on retirement age will be less than a quarter of what it would be if a truly age-neutral system had been implemented.

The renewed social-security debate in Germany, only a few years after the most recent reform in 1992, focuses on further changes in the benefit structure and applicable retirement ages. Major changes, such as a transition from the current PAYG system to a partially or fully funded system, are not seriously debated within the government. Given the large discrepancy between the rate of return of the PAYG system and capital-market returns, a further erosion of the tax base appears likely. The withdrawal of most of those workers who were permitted to opt out appears to be indicative of this.

The window of opportunity for a reform is rather narrow. Population aging shifts the majority of voters even further toward those who benefit the most from a generous PAYG system. In about 10 years, the median voter in Germany will be 50 years old. In addition, considerable time is needed to phase in a reform, due to grandfathering—particularly so, if a partially funded pension system is introduced that requires the accumulation of sufficient savings.

REFERENCES

- Börsch-Supan, Axel. "Population Aging, Social Security Design, and Early Retirement." *Journal of Institutional and Theoretical Economics*, December 1992, 148(4), pp. 583–57.
- . "Germany: A Social Security System on the Verge of Collapse," in Horst Siebert, ed., *Redesigning social security*. Tübingen: Mohr, 1998 (forthcoming).
- Börsch-Supan, Axel and Schmidt, Peter. "Early Retirement in East and West Germany," in Regina Riphahn, Dennis Snower, and Klaus Zimmermann, eds., *Employment policy in the transition: Lessons from German integration*. Heidelberg: Springer-Verlag, 1998 (forthcoming).
- Börsch-Supan, Axel and Schnabel, Reinhold. "Social Security and Retirement in Germany," in Jonathan Gruber and David Wise, eds., *Social security programs and retirement around the world*. Chicago: University of Chicago Press, 1998 (forthcoming).
- Kotlikoff, Lawrence J. and Wise, David A. "Incentive Effects of Private Pension Plans," in Zvi Bodie, John Shoven, and David A. Wise, eds., *Issues in pension economics*. Chicago: University of Chicago Press, 1987, pp. 283–339.
- Lumsdaine, Robin L.; Stock, James H. and Wise, David A. "Three Models of Retirement: Computational Complexity versus Predictive Validity," in David A. Wise, ed., *Topics in the economics of aging*. Chicago: University of Chicago Press, 1992, pp. 16–60.
- Meghir, Costas and Whitehouse, Edward. "The Job Exit Behaviour of Older Men in the UK: Evidence from Event History Data." Working paper, Economic and Social Research Council, Cambridge, 1993.
- Schnabel, Reinhold. "Internal Rates of Return of the German Pay-As-You-Go Public Pension System." University of Mannheim Discussion Paper SFB 504, 1997.
- Siddiqui, Sikandar. "Labour Supply Disincentive Effects of Old Age Public Pensions: A Case Study for West Germany Combining Panel Data and Aggregate Information." CILE Working Paper No. 28, University of Konstanz, September 1995.
- Stock, James H. and Wise, David A. "The Pension Inducement to Retire: An Option Value Analysis," in David A. Wise, ed., *Issues in the economics of aging*. Chicago: University of Chicago Press, 1990, pp. 205–30.

INFORMING RETIREMENT-SECURITY REFORM[†]

401(k) Plans and Future Patterns of Retirement Saving

By JAMES M. POTERBA, STEVEN F. VENTI, AND DAVID A. WISE*

The recent growth of 401(k) retirement saving plans has significantly affected the composition of personal saving flows in the United States. In 1993, the most recent year for which complete data are available, 23.1 million individuals participated in 401(k) plans. Employer and employee contributions to these plans totaled \$69.3 billion, when the annual flow of personal saving in the national income accounts was \$248.5 billion. Anecdotal evidence suggests that 401(k) plans have grown rapidly in the years since 1993, and that the flow of annual contributions may now be approaching \$100 billion.

The expansion of 401(k) plans will affect the level and composition of individual financial assets at retirement for future generations of retirees. We have argued in previous work, summarized in Poterba et al. (1996), that most 401(k) contributions represent "new saving," and that the growth of the 401(k) system has raised household net worth relative to what it would have been without 401(k) plans. We therefore view 401(k) assets as a significant, and new, source of wealth for future retirees. Even if 401(k) contributions do not represent new saving, however, as Eric Engen et al. (1996) have argued, it is important to track the prospective growth of 401(k) assets. Individuals have much greater discretion in the investment and decumulation of 401(k) assets than in many other retirement

saving programs, such as defined-benefit pension plans. The prospective growth of 401(k) assets indicates the size of the asset pool over which individuals will have such discretion.

In this brief paper we summarize current participation and contribution patterns in 401(k) plans. We track the contribution behavior of specific age cohorts during the last 15 years, as 401(k) plans have expanded, and we use this information to project 401(k) balances at retirement age for workers who are currently between the ages of 30 and 40. Future 401(k) balances will depend on several factors, including the growth rate of individual earnings over the next three decades, the investment-allocation decisions made by 401(k) participants, and the incidence of pre-retirement withdrawals from 401(k) accounts. We conclude by discussing the prospects for continuing growth of 401(k)-type retirement saving plans.

I. Participation and Contribution Behavior in 401(k) Plans

There are two primary sources of information on participation in 401(k) plans. The first is the set of Form 5500 reports that pension plans are required to file with the Department of Labor. These reports provide information on the aggregate number of 401(k) participants and the flow of contributions to these plans, but they do not provide information on the demographic or economic characteristics of plan participants. The second source of 401(k) data is household surveys, such as the Survey of Income and Program Participation (SIPP) and the Current Population Survey (CPS) Employee Benefit Supplement. We rely on CPS data for the tabulations shown below but note that the reported rate of 401(k) participation in the CPS is somewhat higher than that in the SIPP.

[†] Discussants: Olivia Mitchell, University of Pennsylvania; Sylvester J. Schieber, Wyatt Company.

* Department of Economics, Massachusetts Institute of Technology, Cambridge, MA 02142-1347; Department of Economics, Dartmouth College, Hanover, NH 03755; and J. F. Kennedy School of Government, Harvard University, Cambridge, MA 02138, respectively. We are grateful to Olivia Mitchell and Sylvester Schieber for helpful comments and to the National Institute on Aging and the National Science Foundation (Poterba) for research support.

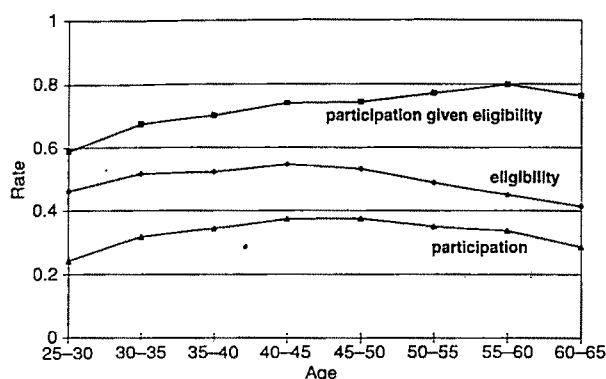


FIGURE 1. 401(k) ELIGIBILITY, PARTICIPATION GIVEN ELIGIBILITY, AND PARTICIPATION RATES BY AGE, 1993 CPS

Figure 1 shows the age profile of 401(k) eligibility, participation given eligibility, and participation in 401(k) plans as indicated in the 1993 CPS. The unit of observation for these tabulations is the family. We define a family as participating in a 401(k) plan if any of the employed individuals in the family participate. The figure shows that participation conditional on eligibility rises gradually with age, while eligibility rises sharply between ages 25 and 45. The net effect of these trends is a modest age-related increase in 401(k) participation rates until age 50.

We are interested in projecting future 401(k) asset-accumulation patterns, and for this purpose, current age-specific 401(k) participation rates may be of limited use. The participation rate of 55-year-olds in 1993 is likely to understate the 401(k) participation rate of current 35-year-olds when they reach age 55. To explore the trends in 401(k) utilization, we track the 401(k) participation rates for several age cohorts using consecutive waves of the CPS. Figure 2 presents the results of our cohort analysis. The rate of 401(k) participation has risen more quickly over time for each cohort than the cross-sectional age-participation pattern at any date would have suggested. This suggests that if the current trends in the growth of 401(k) plans continue, households reaching retirement in the future will have a much higher probability of having participated in a 401(k) plan than those who reached retirement in the early 1990's. For the purposes of our forecasts below, we assume that, when the

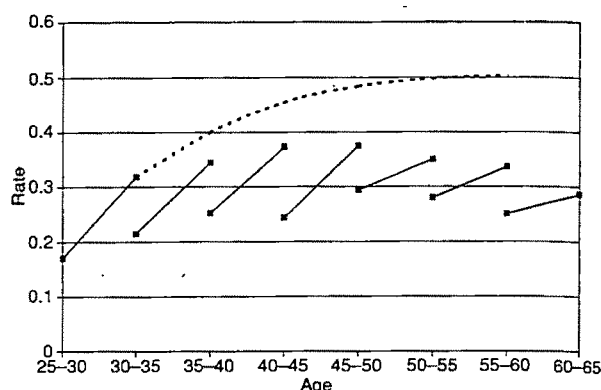


FIGURE 2. 401(k) PARTICIPATION RATES BY COHORT

cohort of households headed by individuals who were 33 years old in 1993 reaches age 55, their 401(k) participation rate will be 50-percent greater than that of 55-year-olds in 1993. The dashed line in Figure 2 shows this participation profile. Given the continued growth of 401(k) plans since 1993, this is probably a conservative assumption.

There are important income-related differences in 401(k) participation and eligibility, but these are not evident in aggregate statistics like those shown in Figures 1 and 2. To document these differences, we stratified families in the 1993 CPS by earnings decile and tabulated 401(k) participation rates for each decile. Table 1 presents our findings. The table shows that the participation rate for those in the top two earnings deciles averages more than 50 percent, while the participation rate in the lowest two deciles averages under 10 percent. The sharp gradient in 401(k) participation by earnings class is the result of substantial differences in 401(k) eligibility rates across earnings classes, as well as a rough doubling of participation rates given eligibility between the lowest and highest earnings deciles. We recognize these differences in our projections of future 401(k) asset accumulation by disaggregating our projections by earnings decile. At present, 401(k) plans are more likely to be available at firms that employ highly paid workers than at firms that employ lower-wage workers. One of the unresolved issues in our analysis, and an important determinant of the future course of 401(k) asset growth, is the extent to which these plans will diffuse to low-wage firms.

TABLE 1—PARTICIPATION IN AND ELIGIBILITY FOR 401(k) PLANS, 1993

Earnings decile	Eligibility	Participation given eligibility	Participation
First (lowest)	0.169	0.357	0.050
Second	0.250	0.498	0.109
Third	0.360	0.590	0.190
Fourth	0.477	0.625	0.272
Fifth	0.500	0.629	0.290
Sixth	0.530	0.690	0.340
Seventh	0.628	0.757	0.450
Eighth	0.665	0.776	0.484
Ninth	0.712	0.808	0.548
Tenth	0.715	0.837	0.580
All	0.501	0.708	0.330

Source: Authors' tabulations based on 1993 Current Population Survey, Employee Benefits Supplement. Families are the unit of observation. Differences in the samples for different columns make it impossible to multiply the first (eligibility) column by the second column to obtain the last column.

Projecting future 401(k) asset balances requires information on the amount contributed to these plans as well as the probability of contributing. CPS data show that, on average, a participating family in 1993 contributed a total of 8.7 percent of salary to the plan. Roughly two-thirds of this represented an employee contribution, while the remainder (2.7 percent) was an employer contribution. There is relatively little variation across age or income deciles in the share of salary contributed to these plans. The average share of salary contributed to 401(k) plans, for those who participate in these plans, is substantially greater than the share of earnings that employers contributed on behalf of workers in traditional defined-benefit pension plans. Thus it should not be surprising if 401(k) participants accumulate asset balances that are large relative to defined-benefit pension assets.

Another important input to 401(k) asset projections is the asset allocation chosen by 401(k) plan participants. Historically, investments in corporate stock have yielded substantially higher returns than investments in fixed-income securities such as corporate or government bonds. Ibbotson Associates (1997) report that the arithmetic mean return

on large company stocks over the 1926–1996 period was 12.7 percent, compared with 6.0 percent for long-term corporate bonds and 3.8 percent for Treasury bills. The mean annual inflation rate over this period was 3.2 percent. Poterba and Wise (1998) present data suggesting that between 40 percent and 60 percent of 401(k) assets in 1995 were invested in corporate equities, with a substantial fraction of these assets in the stock of the company that sponsors the 401(k) plan. There is also some evidence that the fraction of 401(k) assets invested in corporate stock has increased over time. In projecting future 401(k) assets we consider three different investment scenarios: a 100-percent corporate bonds portfolio strategy, a 50–50 bonds/equity strategy, and a 100-percent equity investment strategy.

II. Projecting 401(k) Balances for Future Retirees

We estimate the wealth at retirement for future cohorts of retirees by extrapolating the cohort-specific patterns of 401(k) participation shown in Figure 2. We illustrate our findings by focusing on a family that was headed by someone who was 33 years old in 1993 (the date of the most recent CPS information) and who will reach age 65, which we consider retirement age, in 2025. We first estimate the cross-section profile of participation rates by age and earnings decile by fitting a probit model to data from the 1988 and 1993 CPS. The probability of 401(k) participation is modeled as a function of the level and square of the family head's age, as well as a set of ten indicator variables corresponding to earnings deciles. We then adjust the probabilities so that the participation rate for a person of age 55 in 2015 is 50-percent greater than that for a person of age 55 in 1993. (This is the adjustment shown in Fig. 2.) We further assume that anyone participating in a 401(k) plan contributes 9 percent of earnings to the plan.

A critical determinant of 401(k) contributions over an individual's working lifetime is the individual's pattern of labor earnings. While data sets such as SIPP or the CPS provide snapshot information on the earnings distribution, they do not include long earnings histories for respondents. To obtain such

TABLE 2—MEAN WEALTH HOLDINGS (\$1,000's)
BY EARNINGS DECILE, 1992 HRS

Earnings decile	Predicted 401(k) assets	Actual 401(k) assets	Employer pension assets	Nonretirement, non-Social Security wealth	Social Security wealth
First (lowest)	0.3	0.6	39.2	159.9	61.5
Second	0.9	1.0	40.0	103.3	74.1
Third	2.3	2.6	34.4	122.8	84.1
Fourth	3.7	2.2	36.7	129.1	93.4
Fifth	5.4	4.0	52.5	126.5	101.6
Sixth	8.5	6.4	75.7	172.9	108.1
Seventh	12.1	11.3	94.4	175.5	114.7
Eighth	17.6	13.5	105.4	208.7	125.1
Ninth	25.4	19.8	133.1	280.9	132.0
Tenth	37.2	48.7	219.1	535.8	143.4
All	11.4	10.8	82.2	199.7	103.4

Source: Authors' tabulations from the 1992 Health and Retirement Survey. Projected 401(k) assets assume that 401(k) balances are invested 50 percent in stocks and 50 percent in bonds.

information we used the Health and Retirement Survey (HRS), a survey of individuals who were between the ages of 51 and 61 in 1992. Survey responses to the HRS have been matched to respondents' Social Security earnings records, so it is possible to track earnings profiles for these individuals. A number of data issues forced us to exclude some families from the HRS sample. The most serious restriction was the requirement that both family members had to have complete Social Security earnings histories. The final sample includes 3,992 of the 7,607 original HRS families. Further details of our probit estimates for 401(k) participation, and our selection algorithm, may be found in Poterba et al. (1997).

The HRS asked respondents about their current 401(k) asset balances and various other components of their net worth. Before trying to forecast the 401(k) balances of future retirees, we therefore tried to predict the 401(k) balances in 1992 of HRS respondents, given our historical information on their earnings. The first two columns of Table 2 present the results of this comparison, disaggregated by 1992 earnings decile. Both predicted and actual 401(k) assets rise sharply with earnings. For those in the two lowest earnings deciles, predicted and actual balances are less than \$1,000, while for those in the two highest deciles, predicted balances average \$31,300. Actual balances for those in the two highest deciles average \$34,300. For the entire sam-

ple, the average predicted 401(k) balance is approximately \$600 higher than the actual average balance of \$10,800. This provides encouraging evidence on the predictive accuracy of our 401(k) imputation algorithm.

Table 2 also provides summary information on other components of household net worth for HRS respondents, to place current and future 401(k) balances in perspective. Several patterns are evident in the table. First, nonretirement, non-Social Security assets, which consist of an owner-occupied home and a small amount of non-401(k), non-IRA financial assets for most households, vary between \$100,000 and \$200,000 for almost 80 percent of the population. This confirms the limited role of personal nonretirement saving, noted in Poterba et al. (1994) and many other places, in most households' financial preparation for retirement.

Second, the average household headed by someone between the ages of 51 and 61 in 1992 could expect to receive Social Security benefits with a present discounted value of \$103,400, using the discount-rate assumption in the "Intermediate Assumptions" of the Social Security Board of Trustees. Social Security benefits rise with a household's earnings history, but they vary less with earnings than other components of household net worth. The ratio of the value of Social Security benefits for those in the lowest earnings decile to the value for those in the highest decile is 2.3. The analogous ratio for employer-provided pensions is 5.6. Finally, we note that the current value of 401(k) assets is approximately 10 percent of the value of Social Security wealth. Comparing 401(k) assets to Social Security is a useful metric for evaluating the future growth of these retirement assets.

Table 3 presents our central findings with respect to the future value of 401(k) assets. We use the actual earnings histories of the 1992 HRS respondents, coupled with our projections of future 401(k) eligibility and contribution rates, to make these projections. Assuming that 401(k) assets are invested half in stocks and half in bonds, and that the average returns on these assets over the next 28 years are equal to their average returns over the 1926–1996 period, the average value of 401(k) assets for retirees in 2025 equals

TABLE 3—PROJECTED 401(k) BALANCES AT AGE 65,
COHORT TURNING AGE 65 IN 2025

Earnings decile	All-bond portfolio	50-50 bonds and stocks	All-stock portfolio
First (Lowest)	1.0	1.8	3.6
Second	5.4	10.0	20.0
Third	11.9	22.2	44.4
Fourth	21.7	40.3	80.8
Fifth	28.5	52.5	104.3
Sixth	38.7	71.1	141.3
Seventh	59.8	110.7	221.5
Eighth	77.7	143.2	286.0
Ninth	102.6	187.9	373.2
Tenth	153.9	276.4	540.5
All	50.1	91.6	181.6

Note: All entries are measured in thousands of \$1992, and can therefore be compared to entries in Table 2.

\$91,600. Since the earnings histories of the respondents in our projection are the same as those for the respondents whose net worth was summarized in Table 2, we can compare our projected 401(k) assets with other components of the household balance sheet. In particular, the average value of 401(k) assets in 2025 is approximately 90 percent of the present discounted value of Social Security benefits. If we assumed that all 401(k) assets were invested in equities, the projected average for 401(k) assets would exceed the average value of Social Security benefits by nearly 80 percent.

Table 3 shows that there are substantial differences across earnings deciles in projected 401(k) balances. For the lowest earnings decile, we project only \$1,800 in 401(k) assets. For households in the center of the earnings distribution, projected 401(k) assets average nearly \$60,000. For households in the top earnings decile, projected 401(k) balances are \$276,400.

III. Conclusions and Implications

The projections in Table 3 demonstrate that 401(k) plans are likely to play a central role in providing for the retirement income of future retirees. Younger workers are more likely to participate in 401(k) plans than are older workers, and younger firms are less

likely to offer traditional defined-benefit pension plans than are older firms. The net effect of these trends will be an important shift in the future composition of retirement-income provision.

Whether our projections are realized will depend on several factors. One is the future expansion of 401(k) plans at small firms where 401(k) availability has historically been lower than at large firms. Our projections are largely based on the assumption that current participation rates of young cohorts do not decline as those cohorts age, but if current 401(k) expansion trends slacken, then our projections may overstate future 401(k) balances. A second key factor in our projections is the allocation of 401(k) assets between stocks and bonds. If current trends toward greater investment in corporate equities continue, and if equity returns match their historical averages, our projections could substantially understate future 401(k) balances. Finally, an important factor in future 401(k) wealth is the degree to which workers draw down 401(k) balances before they reach retirement. Our algorithm for predicting 401(k) assets at retirement does not allow for preretirement withdrawals. Poterba et al. (1998) and the Employee Benefit Research Institute (1997) present data suggesting that most of the dollars that are distributed from 401(k)-type plans are reinvested in other retirement saving plans. However, the incidence of preretirement withdrawals may increase in the future, reducing the accumulating value of 401(k) plan assets.

Whether growing 401(k) plan balances will provide a supplement to the principal current sources of retirement income, Social Security and employer-provided pensions, or will partly make up for reductions in these other sources of financial support is an open question. Many proposals currently under discussion would reduce the role of government transfers in providing for the health and financial support of elderly households; others would shift responsibility for retirement income and health insurance from firms to workers. If these proposals are enacted, the growing stock of 401(k) wealth could provide an important source of replacement assets for affected households.

REFERENCES

- Employee Benefit Research Institute.** "Good News for Retirement Income Security: Lump-Sum Distribution Rollovers on the Rise." Washington: EBRI, 1997.
- Engen, Eric; Gale, William and Scholz, J. Karl.** "The Illusory Effects of Saving Incentives on Saving." *Journal of Economic Perspectives*, Fall 1996, 10(4), pp. 113-38.
- Ibbotson Associates.** *Stocks, bonds, bills, and inflation: 1997 yearbook*. Chicago: Ibbotson Associates, 1997.
- Poterba, James M.; Venti, Steven F. and Wise, David A.** "Targeted Retirement Saving and the Net Worth of Elderly Americans." *American Economic Review*, May 1994 (*Papers and Proceedings*), 84(2), pp. 180-85.
- _____. "How Retirement Saving Programs Increase Saving." *Journal of Economic Perspectives*, Fall 1996, 10(4), pp. 91-112.
- _____. "Implications of Rising Personal Retirement Saving." National Bureau of Economic Research (Cambridge, MA) Working Paper No. 6295, 1997.
- _____. "Lump Sum Distributions from Retirement Saving Plans: Receipt and Utilization," in David A. Wise, ed., *Inquiries in the economics of aging*. Chicago: University of Chicago Press, 1998, pp. 85-105.
- Poterba, James M. and Wise, David A.** "Individual Financial Decisions in Retirement Savings Plans and the Provision of Resources for Retirement," in Martin Feldstein, ed., *Privatizing social security*. Chicago: University of Chicago Press, 1998 (forthcoming).

The Cause of Wealth Dispersion at Retirement: Choice or Chance?

By STEVEN F. VENTI AND DAVID A. WISE*

In the United States, some households reach retirement having accumulated substantial wealth while others approach retirement with very limited assets. Considerable research has been directed to testing alternative explanations of why people save. We ask a different question: how much of the dispersion in wealth at retirement arises because of variation in resources out of which households could have saved over their lifetimes, and how much of the dispersion arises because some households saved from available resources while others did not save?

Whether differences in accumulated wealth reflect the availability of financial resources or thrift can have important policy implications. In the United States, income from both capital and labor is taxed at the federal level, and wealth is often taxed at the state and local level via the property tax. A number of other features of the tax system either directly or indirectly impose *ex post* taxes on accumulated wealth. Examples include means-tested taxes on Social Security benefits, implicit taxes on college financial-aid programs, "success" tax penalties on excess accumulations of pension saving (now abolished), estate taxes, and spend-down provisions in Medicaid. Thus two households with the same lifetime earned income pay very different levels of lifetime taxes if they save different proportions of their earnings. The tax difference consists of levies on capital income and accumulated wealth.

The legislative choice of such taxes may be favored by a perception that wealth dispersion arises principally from "chance" events such as inheritances, poor health, or other "shocks" to wealth that limit the resources out of which saving could be taken. In this case,

such taxes may be considered "fair" because they work to equalize wealth differences arising because of luck. On the other hand, if the dispersion of wealth among the elderly reflects conscious spending versus saving decisions while young, so that saving varies dramatically across households with similar lifetime earned incomes, taxes on accumulated saving may be harder to justify and appear to penalize savers who spend less when they are young. In addition, from an economic perspective, taxing saving has no incentive effects if wealth accumulation is random. But this is not true if wealth accumulation results from conscious decisions to save versus spend while young. In this case, levies on savers may have substantial incentive effects, discouraging individuals from saving for their own retirement. This paper is about the reasons for the dispersion in the accumulation of assets of persons with similar lifetime earnings. This is not a new issue. In 1953, Milton Friedman in "Choice, Chance, and the Personal Distribution of Income" concluded: "Yet I think it [the analysis] goes far enough to demonstrate that one cannot rule out the possibility that a large part of the existing inequality of wealth can be regarded as produced by men to satisfy their tastes and preferences."

The analysis is based on the baseline interview of the Health and Retirement Survey (HRS) which surveyed households with a person aged 51-61 in 1992. The survey includes a complete accounting of assets, including personal retirement assets such as IRA's and 401(k) balances, other personal financial assets, home equity, and assets such as real estate and business equity. Employer-provided pension wealth is calculated from the respondent's description of provisions of the pension plan. Perhaps the most important advantage of this survey is that it has been matched to earnings histories provided by the Social Security Administration. The Social Security earnings histories, together with linked Current Population

* Department of Economics, 6106 Rockefeller Center, Dartmouth College, Hanover, NH 03755, and National Bureau of Economic Research, 1050 Massachusetts Avenue, Cambridge, MA 02138-5398, respectively.

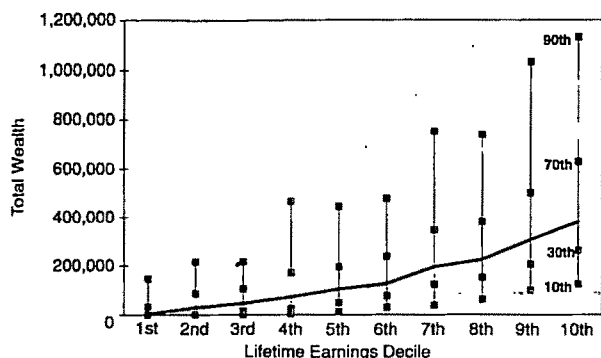


FIGURE 1. THE 10TH, 30TH, 50TH, 70TH, AND 90TH PERCENTILES OF TOTAL WEALTH BY LIFETIME EARNINGS DECILE

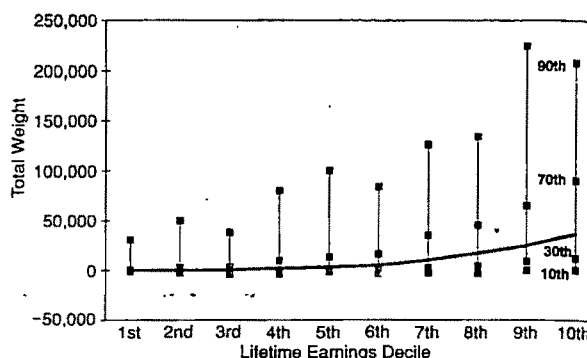


FIGURE 2. THE 10TH, 30TH, 50TH, 70TH, AND 90TH PERCENTILES OF FINANCIAL ASSETS BY LIFETIME EARNINGS DECILE

Survey data, are used to determine lifetime earned income. All of the analysis is based on household rather than individual data. In all, 3,992 households are used. Details are available in Venti and Wise (1997).

The next section of the paper documents the enormous variation in wealth on the eve of retirement. We find substantial variation in wealth even after controlling for lifetime earned income. Next we consider the link between wealth and lifetime income. We then set aside the dispersion attributable to differences in lifetime earned income and ask whether the wealth dispersion among families with similar lifetime earnings can be explained by individual circumstances that, over the course of a lifetime, may have limited the household's ability to save for retirement. We also consider how much of the dispersion can be explained by different investment choices made by households. Dispersion in wealth not accounted for by lifetime financial resources, circumstances affecting ability to save, or investment choices is interpreted as a broad indicator of the dispersion in the "taste" for saving among households. A final section concludes.

I. How Dispersed Is Wealth?

Perhaps most striking is the enormous dispersion of wealth within lifetime earnings deciles, as shown in Figure 1. For example, in the fifth lifetime income decile, the 90th quantile is 35 times as large as the 10th quan-

tile. The range is less extreme in higher earnings deciles but is still very wide: factors of 16, 19, 12, 10, and 9 in the sixth through the tenth lifetime earnings deciles, respectively. It is evident that many households with the highest lifetime earnings have accumulated very little wealth, while some with the lowest lifetime earnings have accumulated a great deal. For example, the 90th quantile is approximately \$150,000 for the lowest decile and is well above \$200,000 for the second and third deciles.

Figure 2 shows the dispersion of financial assets (excluding personal retirement assets such as IRA and 401(k) accounts). That most people do not save much is not new. That many of those with high earnings save so little is, however, striking. The 10th quantile is negative or close to zero for every lifetime earnings decile! The same is true for the 20th quantile, with the exception of the highest earnings decile for which the 20th quantile is a paltry \$6,400. The medians range from zero for the lowest three deciles, to \$3,000 and \$5,800 for the fifth and sixth quantiles, to \$10,000 for the 70th to \$36,500 for the highest earnings decile. Like the dispersion in total wealth, the range of personal financial assets from the 10th to the 90th quantiles is extremely broad.

II. Wealth and Lifetime Earnings

Households with higher earnings of course save more than those with low earnings. But do they save a larger proportion of

TABLE 1—MEAN AND MEDIAN
WEALTH-TO-EARNINGS RATIOS

Lifetime earnings decile	Mean wealth/earnings	Median wealth/earnings
First	1.76	0.60
Second	0.65	0.39
Third	0.45	0.32
Fourth	0.45	0.29
Fifth	0.41	0.27
Sixth	0.40	0.28
Seventh	0.42	0.29
Eighth	0.40	0.29
Ninth	0.51	0.34
Tenth	0.45	0.36

earnings? Table 1 suggests not. It shows the ratio of mean lifetime earnings to mean wealth at retirement (and median lifetime earnings to median wealth) by earnings decile. Although the estimates for the bottom and top lifetime earnings deciles may be imprecise, the data show that the ratio of wealth to lifetime earnings is quite flat over the central range of lifetime-earnings deciles. The ratio declines in the lower intervals and rises at the upper tail. Thus it appears that at this level of aggregation there exists a more or less linear relationship between the lifetime earnings and the ratio of wealth to lifetime earnings.¹

III. Wealth Dispersion Given Lifetime Earnings

Figure 1 demonstrates that there is substantial variation in wealth between lifetime earnings deciles: the range in median wealth between the first and tenth deciles is \$377,000. But the figure also shows enormous variation within deciles. Using a regression framework, we find that the unconditional standard deviation in wealth is reduced by only 5.05 percent when we control for the decile of lifetime earned income.

Thus primary attention in this paper is given to the reasons for the dispersion in

wealth within lifetime earnings deciles. We want to distinguish among several possible explanations for the dispersion. The first is the set of household circumstances that may have limited or enhanced the financial resources available for saving, given lifetime earnings. The second is the set of investment choices made by those who save. The third, that which is not accounted for by the first two, is the choice to save out of available resources.

The approach we follow is to determine how much of the variation in wealth can be explained by household attributes associated with the first two explanations. Within each lifetime earned-income decile we compare the distribution of wealth adjusted for each set of factors to the unconditional distribution. The exposition is necessarily graphical for the most part. More standard measures of the reduction in variance are presented in Venti and Wise (1997).

The first set of attributes consists of individual circumstances that may enhance or limit resources that may be available for saving. We think of them as "resource shocks" or "chance" events. These include inheritances, gifts, and health status. We also control for age, number of children, and marital status. Poor health may reduce lifetime earnings, but given earnings may also limit the resources out of which saving could be taken. We have only limited information on lifetime health and use health status at the time of the survey as an indicator of longer-term health status. Although children and marital status are not literally chance events, we include them because it is likely that expenses associated with raising children also reduce the pool of resources that could be saved. In particular, given the age range in the sample (51–61 years), it is possible that education expenditures can account for some of the household variation in wealth. Marital status is also included because, if only due to economies of scale, it may determine resources out of which saving could plausibly be drawn. Finally, age is included because the range of ages of HRS household heads is likely to be systematically related to asset accumulation. We do not include education,

¹ The question of whether higher-income households have higher saving rates has a long history in economics. Karen E. Dynan et al. (1996) review the evidence.

ethnic group, and other attributes that may be correlates of the "taste" for saving.

Within each lifetime earnings decile, we first predict wealth with a simple specification of the form

(1) Wealth

$$\begin{aligned}
 &= \text{Constant} + \beta_1(\text{Married}) \\
 &+ \beta_2(\text{Never Married}) \\
 &+ \beta_3(\text{Widowed, Divorced,} \\
 &\quad \text{or Separated}) \\
 &+ \beta_4(\text{No Children}) \\
 &+ \beta_5(\text{Number of Children if } > 0) \\
 &+ \beta_6(\text{Age}) \\
 &+ \beta_7(\text{Poor Health, Single Person}) \\
 &+ \beta_8(\text{Poor Health, 1 of 2 in Family}) \\
 &+ \beta_9(\text{Poor Health, 2 of 2 in family}) \\
 &+ \beta_{10}(\text{No Inheritances}) \\
 &+ \beta_{11}(\text{Amount of Inheritances} \\
 &\quad \text{Received before 1980}) \\
 &+ \beta_{12}(\text{Amount of Inheritances} \\
 &\quad \text{Received 1980 to 1988}) \\
 &+ \beta_{13}(\text{Amount of Inheritances} \\
 &\quad \text{Received after 1988})
 \end{aligned}$$

with appropriate normalizing restrictions for the indicator variables. From this equation, we obtain predicted wealth. Then, within each earnings decile, adjusted wealth is determined by

(2) Adjusted Wealth

$$\begin{aligned}
 &= (\text{Unadjusted Wealth}) \\
 &- (\text{Predicted Wealth}) \\
 &+ (\text{Mean of Wealth})
 \end{aligned}$$

which gives distributions of adjusted and unadjusted (observed) wealth with the same means within each lifetime earnings decile. Note that, if the explanatory factors do not account for the dispersion in wealth, then adjusted and unadjusted wealth will be the same at all points of the wealth distribution within an earnings decile. If the explanatory factors account for a great deal of the dispersion, then the distribution of adjusted wealth will be flat, equal to the mean level of wealth for all households within an earnings decile.

We follow a similar procedure to determine the effect of investment choice on wealth dispersion. Even among households that save the same proportion of earnings, accumulated wealth may differ because some households have invested their savings in the stock market, for example, while others have saved through bank saving accounts or money-market funds. Over long periods of time the average rate of return on stock investments is much higher than the rate of return on money-market funds, although the risk associated with stock investments is also higher. Other households have invested much of their wealth in housing, which will have yielded yet another rate of return. We do not know the investment choices that households made over their lifetimes. The HRS did, however, obtain information on the percentage allocation of financial asset saving (excluding IRA and 401(k) accounts) for five components of financial assets. We use this information, together with information on the proportion of wealth in housing and five other nonfinancial asset categories, as an indicator of the lifetime investment choices of a household. Within each lifetime earnings decile, we again predict wealth, but based on investment choices, with a specification of the form

(3) Wealth

= Constant

+ β_1 (Percentage Wealth in
Personal Financial Assets)+ β_2 (Percentage Financial
Assets in Stocks)+ β_3 (Percentage Financial
Assets in Bonds)+ β_4 (Percentage Financial
Assets in Money-
Market Accounts)+ β_5 (Percentage Wealth
in IRA, 401(k),
and Keogh Accounts)+ β_6 (Percentage Wealth
in Employer Pensions)+ β_7 (Percentage Wealth
in Business Equity)+ β_8 (Percentage Wealth
in Vehicles)+ β_9 (Percentage
Wealth in Housing)+ β_{10} (Percentage Wealth
in Other Real Estate).

We could of course adjust for both factors affecting financial resources and investment choice at the same time. Making separate adjustments to the same base, however, allows us to compare the effect of resource shocks on wealth dispersion with the effect of investment choices on dispersion. The two sets of variables may be correlated, however. To the extent that they are (positively correlated), some

of what is attributed to financial resources in the first adjustment should be attributed to investment choice instead, and some of what is attributed to investment choice in the second adjustment should be attributed to resource constraints. Thus, this procedure maximizes the adjustment attributed to each. (Standard measures of reduction in dispersion, shown in Venti and Wise [1997], suggest that the correlation between the two sets of variables is positive.)

Figures 3A–D show wealth quantiles and compare the distribution of observed wealth to the distribution of wealth adjusted for resource shocks for four of the lifetime-earnings deciles. For each earnings decile, the white bars show the quantiles of adjusted wealth, and the dark bars show unadjusted (observed) wealth quantiles. Overall, the adjustment for circumstances affecting resources available for saving does not have much effect on the dispersion of wealth. For most earnings deciles the adjustment has the largest effect near the top of the wealth distribution. The more detailed analysis in Venti and Wise (1997) reveals that the adjustment reduces the 95th and 98th quantiles in almost every earnings decile, and the reduction in the 98th quantile is especially noticeable. Yet, overall, resource shocks account for very little of the dispersion of wealth within lifetime earnings deciles. At most points of the distribution the effect of these factors is minimal. Figures 4A–D show the results of a similar adjustment for investment choice. As with resource shocks, the effect on wealth dispersion is small. Again, the largest effects appear in the upper tail of the distribution where the adjustment produces some leveling. Resource shocks and investment choices together, when interacted with lifetime earnings, reduce the standard deviation of wealth by about 17 percent.

Thus neither shocks to resources nor investment choices explain much of wealth dispersion within earnings deciles. The controls for household shocks and investment choices are of course imperfect. For example, measures of health at older ages cannot completely capture the series of health shocks that may affect the accumulation of wealth over a lifetime. The same may be true of realizations on equities for the small proportion of

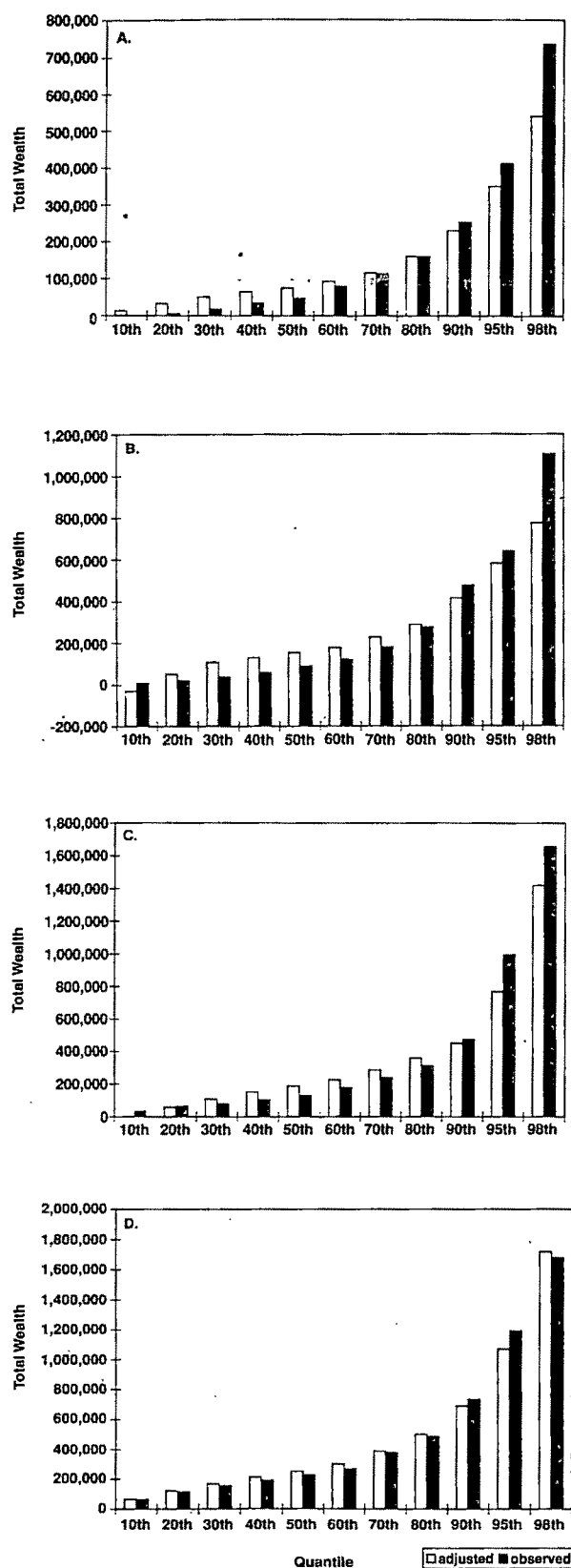


FIGURE 3. WEALTH, OBSERVED VERSUS ADJUSTED FOR "CHANCE" VARIABLES: (A) SECOND LIFETIME-EARNINGS DECILE; (B) FOURTH LIFETIME-EARNINGS DECILE; (C) SIXTH LIFETIME-EARNINGS DECILE; (D) EIGHTH LIFETIME-EARNINGS DECILE

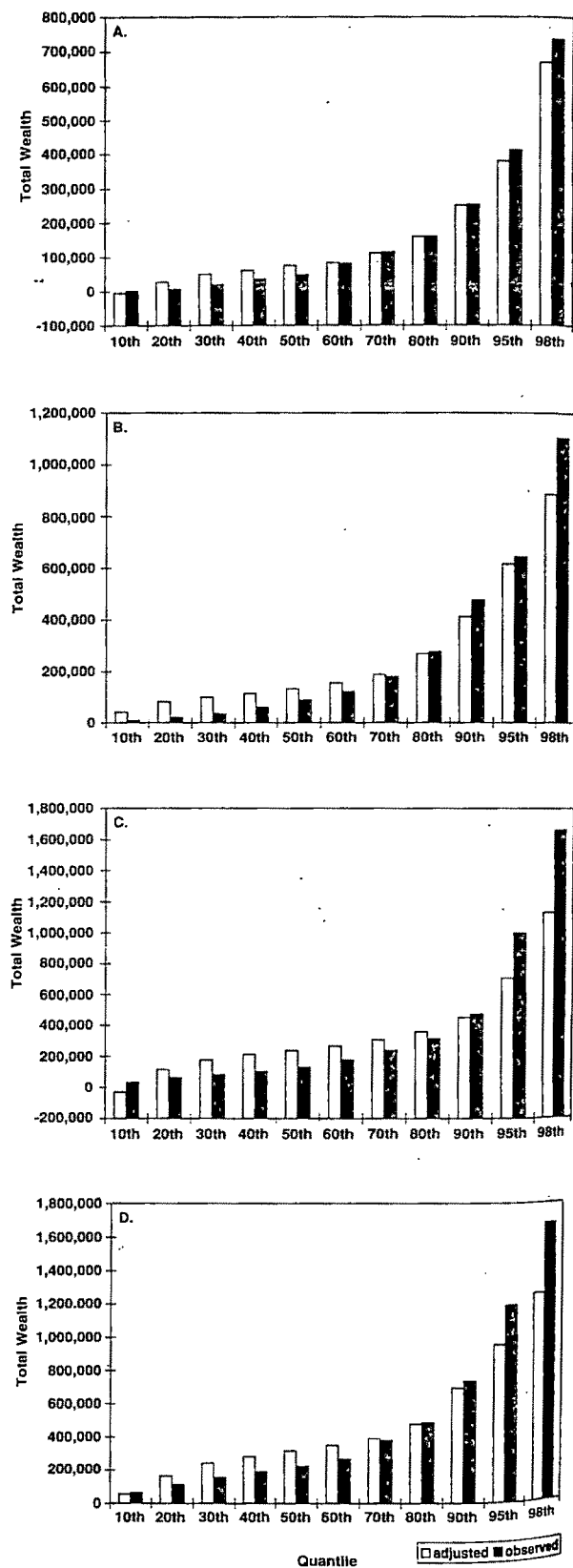


FIGURE 4. WEALTH, OBSERVED VERSUS ADJUSTED FOR INVESTMENT CHOICE: (A) SECOND LIFETIME-EARNINGS DECILE; (B) FOURTH LIFETIME-EARNINGS DECILE; (C) SIXTH LIFETIME-EARNINGS DECILE; (D) EIGHTH LIFETIME-EARNINGS DECILE

households that own stocks. Of course, except for inheritances, persons who do not save will not accumulate wealth under any series of shocks. It seems evident that most of the dispersion must be attributed to the decision to save out of available resources. Some chose to save while young while others chose to spend more and save little. (Controlling for education and ethnic group has only a very modest effect on the wealth distributions within lifetime-earnings deciles, and the results are not reported here.)

In summary, at all levels of lifetime earnings there is an enormous dispersion in the accumulated wealth of families approaching retirement. In the United States it is not only households with low incomes that save little. A significant proportion of high-income households also save very little; and not all low-income households are nonsavers. Indeed a substantial proportion of low-income households save a great deal. Within income deciles, very little of the dispersion in wealth can be

explained by household attributes that may have limited resources available for saving. Nor is much of the dispersion explained by investment choices. Thus the primary determinant of the dispersion of wealth at retirement is evidently the choice to save or spend while young. We find no evidence that chance plays a large role.

REFERENCES

- Dynan, Karen E.; Skinner, Jonathan and Zeldes, Stephen P. "Do the Rich Save More?" Mimeo, Dartmouth College, 1996.
- Friedman, Milton. "Choice, Chance, and the Personal Distribution of Income." *Journal of Political Economy*, August 1953, 42(4), pp. 277-90.
- Venti, Steven F. and Wise, David A. "Choice, Chance, and Wealth Dispersion at Retirement." Mimeo, National Bureau of Economic Research, Cambridge, MA, 1997.

Socioeconomic Status and Health

By JAMES P. SMITH*

The quantitatively large association between many measures of socioeconomic status (SES) and a variety of health outcomes appears pervasive over time and across countries at quite different levels of economic development (Evelyn Kitagawa and Philip Hauser, 1973; Richard G. Wilkinson, 1996). But many analytical difficulties exist in trying to understand its meaning, including the complex dimensionality of health status which produces considerable heterogeneity in health outcomes, the two-way interaction between health and economic status, and the separation of anticipated from unanticipated health or economic shocks.

I present here new evidence on these issues using the first three waves of the Health and Retirement Survey (HRS), a representative national sample of 7,702 households (12,652 individuals) containing a person born between 1931 and 1941. The baseline was fielded during 1992–1993 with follow-ups at two-year intervals. HRS collects excellent information on respondents' employment, income, and wealth (F. Thomas Juster and Smith, 1997). Many different aspects of respondents' health are also measured, including self reports of general health status, the prevalence and incidence of chronic conditions, the extent of functional limitations, and out-of-pocket and total health-care expenditures. The sample used here consists of those 10,236 respondents interviewed in each of the first three waves. All nominal values are expressed in October 1997 dollars.

I. Health and Wealth Transitions

There exists a strong positive association between levels of household income or wealth

and self-reported health status. For example, Smith and Raynard Kington (1997b) report that HRS respondents in excellent health have 2.5 times as much household income and five times as much household wealth as respondents in poor health. Table 1A documents that levels and changes in wealth are also correlated with reported changes in health. Respondents whose reported health status was worse in wave 3 generally had much lower levels of baseline household wealth. The average difference in median wealth for each single threshold health change is \$26,176, amounting to two-thirds of average household income.

While the patterns are not as smooth in panel B of Table 1, median wealth changes between waves 1 and 3 are also correlated with health transitions with an average wealth change of almost \$4,400 per health transition. But this table says nothing about the principal direction of any causal pathway. An economic shock, such as a reduction in wealth, may well affect health outcomes over the long term, but the immediacy of the effects in Table 1 draws attention to the pathway from health to economic status.

One explanation for any cross-wave relationship is that respondents with the same self-reported wave-1 health may be quite heterogeneous in their baseline health. If so, any initial heterogeneity becomes partly revealed in subsequent waves, so that those who report a lower health status by waves 2 and 3 were actually less healthy in wave 1 than those who did not report any health change in subsequent waves. If such heterogeneity is important, part of the association of wealth with changing health status may reflect baseline heterogeneity in wealth and health.

Table 1C examines this possibility by listing a baseline index of mean functional limitations by the joint distribution of general health status in HRS waves 1 and 3. A value of zero in this index implies no functional limitations, and a higher score indicates worse function (Kington and Smith, 1997). The

* RAND, 1700 Main Street, Santa Monica, CA 90407. The expert programming assistance of Iva MacLennan and David Rumpel is gratefully acknowledged. This research was supported by grant P01-AG08291 and grant R01-AG12394 from the National Institute of Aging.

TABLE 1—WEALTH AND HEALTH STATUS

Wave-1 health	Wave-3 health				
	Excellent	Very good	Good	Fair	Poor
A. Median Baseline Wealth (\$1,000's):					
Excellent	232	212	126	NA	NA
Very good	176	178	136	102	NA
Good	140	143	109	74	36
Fair	NA	81	76	57	31
Poor	NA	NA	NA	30	24
B. Median Wealth Change (\$1,000's):					
Excellent	31.7	17.2	10.9	NA	NA
Very good	25.2	18.8	12.6	4.8	NA
Good	14.2	17.9	12.3	2.0	0.2
Fair	NA	0.6	2.9	0.2	-1.3
Poor	NA	NA	NA	0.1	0
C. Mean Wave-1 Functional Status:					
Excellent	5	6	6	8	14
Very good	7	8	10	9	11
Good	9	10	12	15	17
Fair	14	20	19	24	28
Poor	20	36	35	39	45

Note: NA = less than 1 percent of total sample.

patterns are striking. Respondents whose reported health status declined between the waves had more functional limitations in wave 1. The range of variation is quite large. Among those in "fair" health in wave 1, baseline scores range from as low as 14 ("excellent" health in wave 3) to as high as 28 ("poor" health in wave 3). For comparison, there is less than a five-point spread between the bottom and top 20 percent of the income distribution. When stratified only by self-reported health status across waves, the remaining variation in respondents' health is a serious analytical problem. This argues that modeling must incorporate the multidimensionality of health status.

Another difficult modeling issue concerns how much of any health change is actually news to the respondent. Economists are now familiar with the conceptual necessity and empirical complexity inherent in separating out new information or "shocks" in time-series income processes. Similarly, based on currently available information about their current health stock and some prognosis about its fu-

TABLE 2—PREVALENCE AND INCIDENCE OF CHRONIC CONDITIONS

Condition	Baseline prevalence	New incidence	
		Waves 1-2	Waves 2-3
Hypertension	34.3	3.7	3.7
Diabetes	8.6	1.5	2.1
Cancer	4.8	1.1	1.5
Lung disease	6.4	1.5	1.4
Heart condition	11.1	2.3	2.8
Stroke	2.1	0.6	0.9
Arthritis	35.4	6.5	6.7

ture, individuals make uncertain projections about their future health states. These health trajectories contain predictable age-related components, surrounded by considerable individual-level heterogeneity, all of which are updated with the realization of new, and often unpleasant, information about one's health.

One hope of isolating new health information lies in the onset of new chronic conditions. While people may anticipate some onset (e.g., smokers may think they will get cancer), the actual realization, and especially its timing, may be unanticipated. Table 2 presents baseline prevalence rates alongside the percentage of new incidence observed between each pair of successive waves. The pattern of disease prevalence is consistent with that revealed in other sources for this age group. While hypertension and arthritis are particularly common conditions, prevalence rates are not trivial, even for more serious ailments. For example, one in nine HRS respondents has a heart condition, and 5 percent have experienced cancer. Given the two-year window between waves, incidence rates for most conditions are relatively small. While most respondents experience only one new condition, more than one in three report at least one new chronic condition since wave 1. Twenty-three percent of HRS respondents had at least one mild new onset, and 12 percent had at least one severe new onset. Severe conditions were defined as cancer, heart condition, stroke, and diseases of the lung.

The four years spanned by the three HRS waves were marked by considerable health activity. During these years, one-third of all

TABLE 3—OUT-OF-POCKET MEDICAL EXPENDITURES BETWEEN WAVES 1 AND 3

Percentile	Medical expenditures (\$)		
	Severe new chronic condition	Mild new chronic condition	No new chronic condition
10	32	49	22
30	793	434	358
50	1,985	1,072	868
70	4,399	2,255	1,833
90	11,659	6,324	4,774
95	17,108	9,489	7,983
98	31,601	18,322	15,452

respondents were hospitalized at least once, and 5 percent of those hospitalized spent at least one month there. Less than half of all hospitalizations were fully insured, so many respondents experienced some out-of-pocket expenses. Added on to these hospital trips were doctor and dentist visits, outpatient surgery, and drugs. Virtually all respondents visited a doctor at least once, with most visits involving some or complete co-pay. Similarly, 16 percent had outpatient surgery, 60 percent of which involved some respondent payment.

One way health shocks can affect wealth accumulation involves medical expenses associated with new health events. To explore this possibility, Table 3 lists distributions of out-of-pocket (OOP) medical expenses separately for those respondents with new chronic conditions. Compared to respondents who had no onset at all, the median increase in OOP expenses for a severe onset was only \$1,117. These incremental OOP costs were significantly higher in the tails (\$7,000 at the 90th percentile and more than \$16,000 at the 98th percentile). While the typical impact of these health events on OOP costs are modest, there are nontrivial probabilities that the impact might be much larger. Risk aversion and attitudes toward uncertainty then become key parameters in modeling behavioral responses to OOP costs.

Additional medical expenses are not the only way health shocks can affect wealth accumulation. Most directly, healthier people may be able to work more, leading to higher earnings. Reduced savings or the depletion of past asset

TABLE 4—ESTIMATED MEAN EFFECTS FROM NEW ONSETS

	Net worth (\$)	Financial assets (\$)	OOP Medical expenses (\$)	Total medical expenses (\$)
Severe Onset				
Wave 2	-20,927	-9,911	2,305	28,916
Wave 3	-22,973	-3,700	2,226	26,825
Mild Onset				
Wave 2	-7,542	-10,228	476	2,312
Wave 3	-17	-1,276	737	4,244

accumulations may be the preferred first step to cope with episodes of poor health. To estimate the impact of the onset of new chronic conditions, a parallel set of models were estimated predicting total and financial wealth accumulation, and total and OOP medical expenses all measured between waves 1 and 3. In light of the discussion above, in addition to a rather standard set of other covariates, all models included variables designed to capture baseline heterogeneity in health. Evaluated at wave 1, these health variables were self-reports of general health status, the extent of functional limitations, the prevalence of chronic conditions, and whether the respondent and spouse had health insurance. New chronic conditions were separated into their severe and mild variants, and separate estimates were obtained for onset between each set of waves.

Table 4 summarizes estimated mean effects on these four outcomes. While there is legitimate concern about respondents' ability to report medical expenses (especially those costs they do not pay), the ordering of estimated impacts on medical costs are reasonable. For example, coefficients on total medical expenses are about nine times larger for severe compared to mild conditions. While severe onsets impose nontrivial total medical costs (above \$25,000), the impact is considerably muted in OOP expenses where even severe onsets have only about a \$2,250 effect. At least in this age group, the availability of private health insurance significantly softens the immediate financial blow from a health shock.

In spite of these muted effects on OOP expenditures, the estimated effects on wealth ac-

accumulation are not trivial. While new-onset coefficients exhibit much more variability associated with the precise model specification, the average reduction in total wealth due to a severe onset is about \$22,000 (8 percent of average baseline wealth). The estimated total-wealth effects associated with a new mild onset are less stable but considerably smaller. Table 4 also tentatively suggests that, while decrements in financial assets can account for the bulk of any adjustments required by the onset of mild conditions, new severe conditions may require adjustments in other parts of the household wealth portfolio. Since increased OOP medical expenses appear to explain only a small part of the reason that new health shocks reduce wealth accumulation, the reasons must lie elsewhere. A plausible explanation is the labor-supply-induced decreases in household income due to these new onsets.

Table 4 indicates that there are quantitatively significant effects of health on at least one measure of SES: household wealth. Earlier work (Smith and Kington, 1997a, b) based on HRS and AHEAD (respondents at least 70 years old) demonstrated that, at least in older populations, the correlation between health and current-period household *income* mostly reflected causation from health to SES rather than the reverse. However, these strong feedbacks from health status to SES do not deny that there may also be a direct influence of SES on health. Good health is an outcome that people desire, and higher wealth or income enables them to purchase more of it. Similarly, a number of risk factors such as smoking and obesity are more prevalent among those in lower SES groups. However, research indicates that health-care utilization and behavioral risk factors can only explain a small part of the observed association of SES and health (Wilkinson, 1996; Smith and Kington, 1997b).

These findings have led some researchers (especially those associated with the British Whitehall studies) to suggest alternative ways through which SES may affect health. One intriguing hypothesis is that societal-level inequality has a direct influence on health outcomes. There exists a highly nonlinear relation between health and such measures of SES as income and wealth (Smith and Kington, 1997a, b), strongest among those with relatively few resources, weakening

among the middle-class, and almost nonexistent among the affluent. While this nonlinearity would itself produce an association between resource inequality and health, the hypothesis goes beyond individual-level nonlinearity. A common theme is that inequality in relative rank raises levels of psychosocial stress which negatively affects endocrine and immunological processes. In developed countries, it is not material deprivation that matters, but the stress associated with being at the bottom end of an unequal social pecking order. A frequent supporting citation is Robert Sapolsky's (1993) study which indicates that low-ranking male baboons have higher levels of glucocorticoids, apparently a reaction to the chronic stress they experience by their low status. Glucocorticoids are steroid hormones released during stress. Chronic elevated levels during prolonged stress may negatively affect individuals' return to normal functioning.

This hypothesis is important because it provides a direct biological rationale for the reasons why SES may have significant long-run impacts on health status. If true, it raises concerns, given the growing levels of income inequality experienced by many countries in the last few decades. However, the strength of the empirical evidence supporting it remains unconvincing, partly due to the frequent failure to control for the nonlinearity in the individual-level association between SES and health or the significant reverse causation between health and SES documented in this paper.

Another modeling complication is that health status at middle and older ages may reflect health at earlier stages of life, even back to childhood (David J. P. Barker, 1992), and may also be correlated across generations. Smith and Kington (1997a) demonstrate that health outcomes at quite old age (70+) are correlated with health attributes of past, concurrent, and future generations of relatives. Whether this correlation reflects shared genetic endowments or the cumulative impact of common social, economic, and geographic endowments remains an unresolved research question. Similarly, Anders Forsdahl's (1997) study of regional differences in Norway indicates a stronger impact of childhood than adult poverty on the prevalence of coronary heart disease. Since

early health outcomes can affect subsequent decisions such as schooling, marriage, and earnings, it is inappropriate simply to use these outcomes to explain individual variation in current health. Instead, it is necessary to model these feedback mechanisms explicitly and to isolate within-period innovations in the stock of health. While such a research agenda is easier said than done, progress on understanding the critically important relation between SES and health requires it.

II. Conclusions

There is unlikely to be a single winner in the continuing dispute regarding the dual pathways between SES and health. The direct influence of SES on health may be strongest during childhood and early adulthood when levels and trajectories of health stocks become established. Moreover, economic "shocks" may dominate health "shocks" among those in their twenties or thirties as levels of lifetime earnings are determined. The dominant causal pathway may then reverse, as health largely affects SES among those age 50 and older. For example, this paper presents new evidence that new health events have quantitatively large effects on wealth accumulation among those in their fifties. After age 40, the big information people receive is not about their changing economic circumstances, but rather about their overall health and its implications for their eventual mortality and ability to function effectively in old age. Studies that ignore the large impacts that health status can have on SES are simply missing a major part of the story.

REFERENCES

- Barker, David J. P.** "Fetal and Infant Origins of Adult Disease—The Womb May Be More Important than the Home." *British Medical Journal*, 17 November 1990, 301(6761), p. 1111.
- Forsdahl, Anders.** "Are Poor Living Conditions in Childhood and Adolescence an Important Risk Factor for Arteriosclerotic Heart Disease?" *British Journal of Preventive Social Medicine*, June 1997, 31(2), pp. 91–95.
- Juster, F. Thomas and Smith, James P.** "Improving the Quality of Economic Data: Lessons from HRS and AHEAD." *Journal of the American Statistical Association*, December 1997, 92(440), pp. 1268–78.
- Kington, Raynard and Smith, James P.** "Socioeconomic Status and Racial and Ethnic Differences in Functional Status Associated with Chronic Diseases." *American Journal of Public Health*, May 1997, 87(5), pp. 805–16.
- Kitagawa, Evelyn and Hauser, Philip.** *Differential mortality in the United States: A study in socioeconomic epidemiology*. Cambridge, MA: Harvard University Press, 1973.
- Sapolsky, Robert.** "Endocrinology Afresco: Psychoendocrine Studies of Wild Baboons." *Recent Progress in Hormone Research*, 1993, 48, pp. 437–68.
- Smith, James P. and Kington, Raynard.** "Demographic and Economic Correlates of Health in Old Age." *Demography*, February 1997a, 34(1), pp. 159–70.
- . "Race, Socioeconomic Status and Health in Late Life," in Linda Martin and Beth Soldo, eds., *Racial and ethnic differences in the health of older Americans*. Washington, DC: National Academy Press, 1997b, pp. 106–62.
- Wilkinson, Richard G.** *Unhealthy societies: The afflictions of inequality*. London: Routledge, 1996.

Extending the Consumption-Tax Treatment of Personal Retirement Saving

By JOHN B. SHOVEN AND DAVID A. WISE*

A large fraction of Americans now rely almost exclusively on Social Security benefits for support after retirement. The aging population of the United States and our increasing life expectancy, however, are undermining the future financial viability of Social Security. The demographic forces have been compounded by the rapid reduction in the labor-force participation of older Americans. Current Social Security benefits are unsustainable, and the retirement support of future elderly is likely to change substantially. The dramatic change is foreshadowed by the rapid expansion of 401(k) plans, as detailed in the paper by James M. Poterba et al. (1998) in this volume. About \$100 billion is now contributed annually to these plans (an increase from essentially zero in the early 1980's and from about \$50 billion in 1990), and no slowdown in their expansion is yet evident. Contributions to 401(k) plans alone now exceed those to conventional employer-provided defined-benefit and defined-contributions plans combined. Today about 50 percent of employees are eligible for a 401(k) plan, and about 70 percent of these people make plan contributions. When 401(k) contributions are combined with contributions to IRA and Keogh plans, personal retirement plans are even larger. The expanded IRA eligibility provided in the Taxpayer Relief Act of 1997 can only increase contributions to personal retirement plans. Indeed, if traditional employer-provided defined-contribution plans are included, then perhaps 80 percent of total pension contributions are placed in plans in which

investment choices are made by individuals and from which future withdrawals will be determined by individual choices. One might say that the privatization of retirement saving is progressing rapidly, perhaps more quickly than any resolution of the debate about partially privatizing Social Security. Universal 401(k) coverage would indeed look much like a partially privatized Social Security system.

Personal retirement assets (including traditional defined-contribution plans) have an important feature not shared by Social Security and defined-benefit plans. They can be transferred to heirs under arrangements which can be especially advantageous. All pension-plan saving is currently accorded consumption-tax treatment for the life of the plan owner and his or her spouse, subject to minimum distribution requirements after the age of 70.5. What this means is that taxes are only paid when distributions are made. The consumption-tax treatment can end upon the deaths of the plan participant and spouse, when a large fraction of the remaining plan assets can be absorbed by estate and income taxes. Prior to 1982, pension-plan assets were not subject to the estate tax at all, and the consumption-tax treatment passed from one generation to the next; the heir paid income taxes as funds were withdrawn from the plan. By 1985, pension assets were subject entirely to estate taxes. Further, the Tax Reform Act of 1986 introduced the excess-distribution and excess-accumulation excise taxes which imposed 15-percent penalty taxes on "excessively large" distributions from pensions and on "excessively large" pension asset accumulations passed to an estate. These so-called "success taxes" were introduced with essentially no public debate and were recognized by few personal-retirement-plan savers during the decade of their existence. We showed in prior papers (Shoven and Wise, 1996, 1997) that these taxes, when combined with income and estate taxes, severely limited the advantages of

* Shoven: Department of Economics and Dean's Office, dean of the School of Humanities and Sciences, Stanford University, Stanford, CA 94305-2070, and National Bureau of Economic Research; Wise: John F. Kennedy School of Government, Harvard University, and National Bureau of Economic Research, 1050 Massachusetts Avenue, Cambridge, MA 02138-5398.

pension saving for people whose retirement accounts might be subject to them. The combined marginal tax rates faced on pension assets in an estate could reach 96.5 percent and more, virtually confiscating additions to pension assets. The Taxpayer Relief Act of 1997 completely repealed the excess-distribution and excess-accumulation taxes. Once again pension-plan saving is the best way to save for retirement for essentially all Americans.

Pension assets can still face extremely high marginal tax rates, however, when they pass through an estate. For example, before the Taxpayer Relief Act of 1997, the combined marginal state and federal tax rate faced on \$1.2 million in pension wealth as part of a \$3 million estate in California could be 91.97 percent (if the heir was in top federal and state income-tax brackets). For the same case, the effective combined marginal tax rate on pensions was lowered to 82.35 percent with the repeal of the excess-accumulation tax. The heirs are clearly better off, but the estate tax and the personal income tax still combine to produce an exceedingly high marginal tax rate.

Rather than passing pension-account balances through an estate, pension assets can be transferred to heirs in a more tax-advantaged form. Plan distributions and income taxes can be "stretched out" by carefully choosing distribution options and designating would-be heirs as beneficiaries. In this way, part of the pre-1982 consumption-tax treatment of pension assets can be restored. In this short paper, we explain how the process works and illustrate its potential advantages.

I. The Payout Options for Individually Controlled Pension Assets

We assume here that the plan owner wants to maximize the tax-deferral advantage of pensions. Obviously, this is not universally true; many retirees will want to assure for themselves the highest possible standard of living while retired and minimize the risk of outliving their assets. Some may choose to purchase life annuities, although private annuity markets suffer from adverse-selection and moral-hazard problems which make this market inefficient. We have in mind persons who have

other assets to partially finance their retirement consumption and who are concerned with the long-term welfare of their children. Thus they want not only to delay the payment of taxes during their own lifetimes, but to extend deferral over the lifetimes of their heirs. The IRS rules for extending the tax deferral of pensions beyond the lifetime of the participant are extraordinarily complicated, and we can only give illustrative examples here.

The plan owner must first choose an advantageous pattern of taxable distributions in retirement, within the constraints imposed by the IRS minimum-distribution requirements, which were first applied to Keogh plans in 1962 and then to all individually controlled pension plans in 1984. The minimum-distribution requirements are discussed in Mark Warshawsky (1998). To qualify for the tax-advantaged status, distributions (usually over a lifetime) must begin no later than April 1 of the year after the owner attains age 70.5. Before the mandatory commencement of distributions, the participant must choose whether to base the minimum distributions on his or her own life expectancy, the joint life expectancy of the husband and wife, or the joint life expectancy of the owner and someone else, perhaps a child or grandchild. To postpone tax payments, a joint life expectancy would typically be chosen. The minimum distribution is the fraction of assets equal to 1 divided by the remaining life expectancy. If the beneficiary is not the spouse, the "minimum-distribution incidental benefit" rule applies. Effectively, the nonspouse beneficiary will be treated as if he or she is at most ten years younger than the participant. For example, if the participant is age 70 the maximum possible joint life expectancy is 26.2 years (the life expectancy of a 70-year-old and a 60-year-old). The IRS requires that life expectancies be determined by the unisex version of the 1983 Individual Annuity Mortality table published by the Society of Actuaries. Why the IRS imposes life expectancies that are much lower than actual longevity and which even if updated would blatantly underestimate the longevity of women is unclear. Given the complexity of other aspects of pension regulations, it is hard to believe that the answer is simplicity.

Before the first distribution, the participant must also make an irrevocable choice between two ways life expectancy can be adjusted as the participant's age. The first is the so-called recalculation method whereby the participant reexamines the life table each year to determine his or her remaining life expectancy (or a joint life expectancy). In general, life expectancies decrease by less than one year per year: for example, at age 72 life expectancy is only about 10 months less than at age 71. The alternative is the term-certain or "one-year-less" rule whereby life expectancy is reduced each year by precisely one year. There are a number of possible variants. For example, the participant could recalculate, and the spouse could use the one-year-less rule—and there are advantages and disadvantages of each from a tax and estate-planning perspective. To delay taxes, however, the recomputation method appears superior because it minimizes distributions and therefore postpones tax payments. It also avoids a serious risk when using the one-year-less method. With that method, anyone who lives beyond the IRS version of life expectancy at age 70.5 (based on obsolete unisex life tables) will be forced to distribute and pay taxes on all remaining plan assets. Under this one-year-less choice, which is slightly simpler, more than half of all pension participants (and their spouses) will outlive their pension-plan distributions. A large fraction of women owners would outlive their plan distributions.

If the participant's spouse is the beneficiary and if minimum payments are based on recalculated joint life expectancies, there are some interesting asymmetries in the law. If the participant dies before the beneficiary, then the surviving beneficiary can roll the pension funds into her or his individual IRA account, and it is treated as a completely new plan. The beneficiary would not have to make any distributions if she is less than age 70.5 and she then can name her children as beneficiaries and calculate the minimum required distribution based on a joint life expectancy with her child, although the child would be subject to the "ten-year-less rule." When the spouse dies, the child or children can take distribution payments based on the actual life expectancy of the eldest beneficiary. In this manner, the distributions from the initial participant's pen-

sion can be made over the lifetimes of two generations. Payouts of even longer duration could be achieved by naming grandchildren as beneficiaries, although then one would have to be conscious of the generation-skipping tax.

If the beneficiary dies before the participant, however, the participant has no choice but to continue payouts under his or her individual life expectancy, and a new child beneficiary can not be chosen. The money cannot be rolled into an IRA and treated as a "fresh start," as when the plan owner dies first. There are not many ways that the tax-deferral benefits can be extended to the next generation in this case, although one possibility is to arrange for remaining pension assets in the participant's estate to be transferred to a charitable remainder trust with the children as beneficiaries. This is permitted as long as the charity is deemed to benefit from at least 10 percent of the present value of the trust; the other 90 percent can be paid out to the children over their lifetimes.

An entirely different set of considerations apply if the pension participant dies before reaching the required distribution date: April 1 of the year following the attainment of age 70.5. It is important to have a plan "designated beneficiary." Without a beneficiary, the pension assets will become part of the estate, and all benefits must be paid out by December 31 of the year of the fifth anniversary of the participant's death. With a designated beneficiary, the benefits can be paid out over the lifetime of the beneficiary. If the beneficiary is the spouse, she or he can defer withdrawals until the participant would have been 70.5; the spouse can also roll over the inherited pension plan into a new IRA. If the beneficiary is a child or someone else other than the spouse, the benefits can be paid out over the entire life expectancy of the designated beneficiary, but they have to begin by December 31 of the year after the year of the participant's death.

II. An Illustration

With such complicated rules, stretching out plan distributions requires expert advice. We emphasize only that very sizable extensions of the tax deferral are possible. The implications are illustrated by this example, taken from Charles W. Collier (1997): A couple sets up

a distribution plan when the husband (the participant) is age 65 and the wife 64. At that time the plan balance is \$1 million, and no further contributions are anticipated. A nominal rate of return of 9.0 percent is assumed. A hybrid joint-life-expectancy withdrawal plan is chosen (the husband recalculates life expectancy each year, and the wife uses the one-year-less method, with advantages which are too complicated to explain here). The first withdrawal is not required until the participant is roughly age 71, when the husband-wife government-imposed joint life expectancy is 22.1 years. Their first distribution must be at least 4.44 percent ($1/22.1$) of the assets in the plan at that time. When the participant dies at age 80, the original \$1 million has grown to about \$2 million, even after the \$1.3 million paid out in required distributions. The 79-year-old widow rolls the \$2 million into a new IRA and names her 45-year-old daughter as the beneficiary. Doing this immediately is important because of the severely adverse tax consequences if both spouses die before the child is named as beneficiary. The daughter will be treated as if she were 69 for the purposes of determining the joint life expectancy to be used in the minimum-distribution requirements. The government table quotes the joint life expectancy of a 79- and 69-year-old as 17.6 years. The widow dies at age 85, and she will have been required to withdraw an additional \$671,000 from her IRA, but the balance will have continued to grow, reaching just over \$2.3 million. The daughter inherits this \$2.3 million IRA. Her remaining life expectancy, which was 37.7 years when her mother's IRA was established, will be set at 32.7 years, and she is required to use the one-year-less method. The daughter cannot roll the inherited pension into a new IRA (and base withdrawals on a joint life expectancy with a new beneficiary); still, she can continue to make ever-increasing withdrawals until the age of 82.7. If she lives to age 82.7 and if the assumed 9.0-percent rate of return materializes, then the total distribution to the daughter will exceed \$14 million. Distributions from the original participant's plan will have continued 37.7 years after his death and will have totaled roughly \$16 million. The time between the participant's first contribution to the plan and the last with-

drawal from it could easily be 90 years. The tax advantages of this arrangement are enormous relative to the estate and income taxes that are imposed if the assets pass through an estate.

We have skipped some extremely important details. A large pension would trigger a substantial estate-tax burden upon the death of the participant's widow. Here we have implicitly assumed that there are other assets available to pay this tax so that the pension account can be left intact. Stretching out the plan distributions is even more advantageous for much smaller pension amounts which would not trigger estate taxes. The advantage of stretching the plan is apparent if one considers the outcome if the mother had not rolled her husband's pension assets into her own IRA, naming her daughter as beneficiary. First, she would have had to withdraw considerably more in the final five years of her life. But, second, and more important, if the mother's estate were the beneficiary rather than the daughter, then the daughter would have had to distribute and pay income taxes on all money in the plan she inherited within five years of the mother's death. If the daughter wanted to use the proceeds to provide a lifetime source of support for herself, then the cost of being forced out of the pension tax shelter in five years is extremely high. With the "stretch-out" plan, the daughter only pays income taxes as the money is distributed over her entire life expectancy. The investments within the plan continue to compound at the full tax-free rate of return. Without such a plan, income tax would have to be paid on the entire inherited amount within five years of the inheritance, and then, since the money could not be put back into a tax deferred plan, the earnings on the fund would also be subject to full personal income taxation.

III. Summary

The above illustration demonstrates that the consumption-tax treatment of pension assets can be extended well past the death of the participant and his or her spouse. The precise outcome of any stretching-out plan will depend on individual circumstances. The large distributions in the illustration de-

pend importantly on the rate of return being larger than the reciprocal of life expectancy, for example. As a practical matter, with current IRS rules, the advantages of the consumption-tax treatment can only be enjoyed with the help of an expert accountant or lawyer.

REFERENCES

- Collier, Charles W. "The Stretch-out IRA and Charitable Gifts: Is There Life After 70 1/2?" Mimeo, Harvard University (distributed at Harvard University presentation), September 1997.
- Shoven, John B. and Wise, David A. "The Taxation of Pensions: A Shelter Can Become a Trap." National Bureau of Economic Research (Cambridge, MA) Working Paper No. 5815, November 1996; in David A. Wise, ed., *Frontiers in the economics of aging*. Chicago: University of Chicago Press (forthcoming).
- . "Keeping Savers from Saving," in David A. Wise, ed., *Facing the age wave*. Stanford, CA: Hoover Institution Press, 1997, pp. 57–87.
- Poterba, James M.; Venti, Steven and Wise, David A. "401(k) Plans and Future Patterns of Retirement Saving." *American Economic Review*, May 1998 (*Papers and Proceedings*), 88(2), pp. 179–84.
- Warshawsky, Mark. "The Optimal Design of the Minimum Distribution Requirements for Pension Plans." Memo, TIAA-CREF, New York, February 1998.

WOMEN AND RETIREMENT ISSUES

Married Women's Retirement Expectations: Do Pensions and Social Security Matter?

By MARJORIE HONIG*

Twenty-five years ago, married women's retirement decisions were strongly influenced by their husbands' health and retirement status, factors determining the value of women's nonmarket time. In contrast, their own economic opportunity set (wages, Social Security entitlements, and employer pension benefits) appeared to have little effect on their decisions to leave the labor force.¹

Married women currently forming expectations regarding retirement differ in important ways from this earlier generation. They have spent more time in the labor force, earned higher wages, and accumulated substantial pension rights, both private and public. They also have lower probabilities of remaining married. In 1970, 82 percent of U.S. women aged 45–54 were married, while 5 percent were divorced. By 1992, only 73 percent were married, while 16 percent were divorced. Thus, husbands' pension and Social Security benefits are less likely to provide economic security in retirement for the current generation of preretirement married women.

Their expectations regarding retirement should reflect these changing conditions. Relative to earlier cohorts, married women's retirement plans should be more strongly

influenced by considerations of their own economic returns from continued employment. While of interest for its labor-supply implications, this issue is a matter of public concern because of growing evidence that divorce has wide-ranging consequences for the economic well-being of postretirement women (William H. Crown et al., 1993).

Findings from the new Health and Retirement Survey indicate that older married women's expectations of working after age 62 are strongly influenced by their expected wage, nonwage compensation such as employer-provided health and disability insurance, and pension income. Expected Social Security entitlements also appear important, although the evidence for their effect is weaker. Like the earlier generation, wives are also influenced by their husbands' plans, suggesting a tendency toward joint retirement.

I. A Simple Model of Retirement Expectations

The Health and Retirement Survey (HRS) is a nationally representative longitudinal survey focusing on ages 51–61, the "preretirement" years. The notion of preretirement suggests a link between an earlier time and the event of retirement that is made explicit in the life-cycle model of labor supply, in which retirement is seen as the realization of plans formed in the distant past. Considerable evidence supports this hypothesis. Expected retirement age, for example, is explained by many of the factors that determine observed age of retirement (Richard E. Barfield and James N. Morgan, 1978; Arden Hall and Terry Johnson, 1980; Lois B. Shaw, 1984).

Stronger evidence of the predicted link between expectations and realizations is provided in studies that explicitly examine the correspon-

* Department of Economics, Hunter College and the Graduate School of the City University of New York, 695 Park Avenue, New York, NY 10021, and the International Longevity Center of the Mount Sinai Medical Center. This research was supported by the National Science Foundation, the National Institute on Aging/University of Michigan, and the International Longevity Center of the Mount Sinai Medical Center. Unusually able research assistance was provided by Anne C. Krill.

¹ Studies using the Social Security Administration's Retirement History Survey (1969–1979), for example, include Sylvia Pozzebun and Olivia S. Mitchell (1989), Michael D. Hurd (1990), and David M. Blau (1993).

dence between retirement plans and actual retirement. Respondents in the Retirement History Study fairly accurately forecast both their retirement age and Social Security benefits, conditional on the information available to them at the time. When actual deviated from expected retirement, differences were found to be correlated with unforeseeable changes in circumstances (Kathryn H. Anderson et al., 1986; B. Douglas Bernheim, 1988, 1989).

In the first wave (1992) of the HRS, workers were asked the chances, on a scale from 0 to 10 where 0 was equal to "absolutely no chance" and 10 implied "absolutely certain," that they would be working full-time after they reached age 62. Rescaled from 0 to 1, responses to this question may be interpreted as subjective probabilities of continuing work beyond what is now the modal retirement age. Framed in this way, the consistency of individual responses may be examined, unlike questions regarding retirement age. Subjective probabilities have been found to be internally consistent, that is, to have conditional probabilities between 0 and 1, and to correspond with observed retirement probabilities (Hurd and Kathleen McGarry, 1994). They also covary both in the cross section and over time with known determinants of realized retirement (Honig, 1994, 1996). In light of this evidence, I assume that married women's responses regarding the probabilities of working full-time after age 62 are consistent indicators of their eventual retirement status at this age (ignoring partial retirement for these purposes).

At some planning age before retirement, a woman forecasts her future earnings, pension and Social Security entitlements, and value of her nonmarket time. She then forms expectations regarding the likelihood that she will be working at age 62, based on forecasts of her potential income should she retire at 62, the income she would receive if she worked another year, and her expected preferences. The greater her expected income at 62, the lower the likelihood that she will expect to continue working after 62, for given preferences. The greater the return from additional employment, the higher will be her expected income, but the higher also the price of leaving work. If the latter consideration dominates, she will expect to continue working past age 62.

I assume that women's expectations are formed within a traditional family framework so that she treats her husband's expected earnings as a form of property income. These earnings may be discounted by her expectation that the marriage will no longer be intact when she reaches age 62.

I include four components of the expected net reward to work at age 62. I measure the wage expected at age 62 by the natural logarithm of the current hourly wage. Because of growing evidence of the importance of health insurance in retirement decisions, I also include coverage under employer-provided health and disability plans. Finally, because continuing coverage reduces the reward to additional work, I include a variable for whether the plan extends coverage after retirement.

If women consider their Social Security and pension wealth, that is, the expected value at age 62 of the stream of future benefits, rather than the benefits available at 62 (which may be zero), the net reward to work would also include expected changes in the net present value of these assets with an additional year's work. The expectational information elicited in the HRS does not include perceived changes in Social Security or pension wealth, nor is it possible to construct reliable estimates from available data. The findings reported below regarding the importance of the age of eligibility for receipt of employer pensions suggest, moreover, that married women consider their "current" (available at age 62) benefits rather than their future benefit stream.

I measure expected income at age 62 by women's forecasts of their annual Social Security benefits, employer pension benefits, and family income. The latter is measured by total family income net of women's earnings. In constructing expected Social Security benefits, one must account for differences in women's understanding of Social Security program rules. In the HRS, women are asked to report benefits at the age they expect to start receiving them, which may not be 62; they are also asked to form expectations about working at age 62, which are dependent on benefits available at that age. I therefore use three alternative measures of expected benefits at age 62, each one reflecting a different understanding of the rules regarding changes in benefits according to age.

I construct expected annual pension benefits at age 62 from detailed information reported by women on the three most important plans in which they are enrolled in their current jobs. These plans may be either "defined benefit" (benefits based on a formula involving age, years of service, and salary), "defined contribution" (benefits based on employer and employee contributions and return on investment), or a combination of the two. Summary information on expected benefits from additional plans is also included.

Among younger married women, the value of nonmarket time is primarily a function of the number of children in the home and tastes for nonmarket activities. For preretirement women, preferences are more likely to be spouse-related and, in the context of retirement expectations, to focus on spouses' anticipated retirement.² I therefore include a measure indicating whether the husband, conditional on having formulated his own expected retirement age, expects to be working when his wife is age 62. I also include a variable indicating whether the wife places a high value on time to be spent with her husband in retirement. These measures capture not only the value of home production arising from household needs related to her spouse, but also the value that she places on joint leisure activities.

II. Data and Findings

My sample of 1,229 white married women aged 51–61 with working spouses is drawn from Wave 1 of the Health and Retirement Survey. It excludes women who were self-employed or working part-time at the interview as well as those for whom critical data were missing or proxy interviews were obtained.

I estimate the subjective probability of working full-time after reaching age 62, which I refer to as the expected retirement function, on the 590 full-time (more than 30 hours per week) wage- and salary-earners in this sample. Since the parameters of this function are esti-

mated within the subsample of workers, they are conditional on the participation decision. The positive and significant coefficient of the inverse Mills ratio, which is included to correct for potential selectivity, indicates that the working women within the larger sample have positive residuals in expected retirement, suggesting stronger tastes for work.

Parameters of the expected retirement function, estimated by maximum likelihood, appear in Table 1. These estimates provide support for the premise that the current generation of married working women take into account their own opportunity set in forming expectations about their retirement. However, their husbands' expected retirement age remains a powerful influence on their plans, as it did in the previous generation.

The natural logarithm of the hourly wage, an estimate of the expected wage at age 62, has a significant and positive effect on the probability of continued full-time work after age 62.³ Evaluated at the sample means, a 10-percent increase in the current wage increases the probability of continuing work by 7 percent. The positive sign of the wage coefficient indicates that a higher expected price of nonmarket time has a stronger effect on the likelihood of continuing work than higher income does.

Wives' own employer-provided health insurance also has a significant effect on their expected chances of continuing work, raising the subjective probability by 30 percent. Continuation of this coverage after retirement, which reduces the expected net reward to work, lowers the subjective probability by 19 percent.⁴ Employer-provided insurance against disability also has a strong effect on the probability of future work, increasing the subjective probability by nearly 20 percent.

Expected income at age 62 includes forecasted Social Security benefits, employer pen-

² The coefficient of a variable measuring number of children in the home was not significantly different from zero, and the variable was not retained.

³ Treated as jointly determined with the subjective probability of working past 62 and corrected for selectivity. Estimates of the wage-determination and participation (probit) equations and variable means are available on request.

⁴ This coefficient is greater than the standard error but below conventional levels of significance.

TABLE 1—PARAMETER ESTIMATES OF EXPECTED RETIREMENT FUNCTION, FULL-TIME WAGE- AND SALARY-EARNERS

Variable	Coefficient	Standard error
ln(hourly wage) ^a	0.120 [†]	0.074
Health insurance on current job	0.106*	0.048
Health insurance after retirement	-0.074	0.055
Disability insurance on current job	0.066*	0.032
Expected annual Social Security benefits at age 62 (\$1,000's) ^b	-0.009	0.007
Eligibility for pension at age 62 ^c	-0.206**	0.054
Expected annual pension benefits at age 62 (\$1,000's) ^d	-0.003	0.004
Family income (excluding wife's earnings; \$1,000's)	-0.001**	0.000
Husband reports expected retirement age	-0.089*	0.036
Husband expects to work when wife is 62	0.098**	0.037
Time with spouse in retirement important	-0.136**	0.030
Inverse Mills ratio	0.084*	0.040
Intercept	0.343	0.196
Log likelihood:	-531.804	

Notes: *N* = 590 white married women aged 51–61 in 1992, spouse employed. The dependent variable is the subjective probability of working full-time after age 62 (scale from 0 = no chance to 1 = absolutely certain); mean = 0.44 (SD = 0.38); estimation method = maximum likelihood (ML).

^a Treated as endogenous and corrected for selectivity.

^b Conditional on knowing benefits.

^c Conditional on knowing age of eligibility.

^d Conditional on eligibility and knowing benefit.

[†] Statistically significant at 10-percent level (two-tailed test).

* Statistically significant at the 5-percent level (two-tailed test).

** Statistically significant at the 1-percent level (two-tailed test).

sion benefits, and family income excluding wives' earnings. The latter has a significant effect on the expected probability of working after 62, although the magnitude is small: a 10-percent increase in income reduces the likelihood by 2 percent. Social Security benefits also have the predicted negative effect on the

expected probability of working. The coefficient⁵ is not precisely estimated but provides weak support for the role of Social Security income in expected retirement. The effect is identical in magnitude to that of family income. Among the three alternative measures of benefits available at age 62, the one assuming the most complete knowledge (of both the actuarial reduction for Social Security benefit receipt before 65 and the delayed retirement credit for first receipt between ages 66 and 69) has the greatest explanatory power.

Employer pensions, in contrast, have a strong effect on women's expectations about working after age 62. Eligibility for benefits at age 62 reduces the subjective probability of continuing work by 27 percent.⁶ Among women eligible for benefits at 62, however, the value of expected benefits does not have a significant influence on probabilities of continued work. The coefficient of the benefits variable has the predicted negative sign but is slightly smaller than the standard error. The point estimate indicates an effect about the same size as that of Social Security benefits and family income. These results suggest that, while married women may not be influenced by marginal changes in pension benefits (or may not accurately forecast their benefits), the expected receipt of pension income is an important determinant of retirement timing. A second implication is that women who will not be eligible for benefits at age 62 do not appear to consider the discounted value at age 62 of their future stream of benefits; that is, they do not consider their pension wealth. Rather, the significance of the eligibility condition implies that they assign a value of zero to expected pension income at age 62. They may do so because they expect borrowing constraints at retirement, have a short life expectancy (although health was introduced and was not

⁵ Conditioned on expected eligibility for benefits and knowledge of their expected value.

⁶ Since eligibility at age 62 is conditioned on coverage and knowledge of age of eligibility, the total effect is measured by the sum of coefficients of the conditioning variables (not reported) and the eligibility indicator. The potential confounding effect of eligibility for Social Security benefits at age 62 is controlled for by the Social Security eligibility variable (not reported).

significant) or are risk-averse or shortsighted. A measure of pension wealth that allowed for nonzero benefits at age 62 for women eligible after age 62 (but could not account for expected changes in wealth due to plan payout structure) was less important than benefits estimated among eligible women and did not reduce the effect of eligibility.

While the findings above suggest that married women today place greater weight on their own opportunities in considering retirement, they share with the previous generation a high regard for shared leisure time in retirement. Husbands' retirement plans are significant determinants of women's plans to continue work after age 62. Wives of men who, having decided on a retirement age, expect to be working when their wives are 62 have a 27-percent higher expected probability of working after age 62. Overall, wives of men who have selected a future retirement age have a 26-percent lower expected probability of working after age 62. This latter finding is consistent with plans for joint retirement if one assumes that husbands' determination of a retirement age indicates intentions to retire in the more immediate future. Finally, women who place a high value on time to be spent with their husbands in retirement have a 37-percent lower expected probability of working after age 62.

III. Conclusion

These findings suggest that older married women today, in forming expectations about when to retire, not only consider their husbands' retirement plans, but also systematically evaluate their own future opportunity sets. As a result, they may be less vulnerable than current retirees to the negative consequences of divorce and diminished spousal support in retirement.

REFERENCES

- Anderson, Kathryn H.; Burkhauser, Richard V. and Quinn, Joseph F. "Do Retirement Dreams Come True? The Effects of Unanticipated Events on Retirement Plans." *Industrial and Labor Relations Review*, July 1986, 39(4), pp. 518-26.
- Barfield, Richard E. and Morgan, James N. "Trends in Planned Early Retirement." *Gerontologist*, August 1978, 18(4), pp. 13-18.
- Bernheim, B. Douglas. "Social Security Benefits: An Empirical Study of Expectations and Realizations," in Rita Ricardo-Campbell and Edward P. Lazear, eds., *Issues in contemporary retirement*. Stanford, CA: Hoover Institution, 1988, pp. 312-50.
- _____. "The Timing of Retirement: A Comparison of Expectations and Realizations," in David A. Wise, ed., *The economics of aging*. Chicago: University of Chicago Press, 1989, pp. 335-58.
- Blau, David M. "Labor Force Dynamics of Older Married Couples." Working Paper, University of North Carolina, 1993.
- Crown, William H.; Mutschler, Phyllis H.; Schulz, James H. and Loew, Rebecca. *The economic status of divorced older women*. Waltham, MA: Policy Center on Aging, Brandeis University, 1993.
- Hall, Arden and Johnson, Terry. "The Determinants of Planned Retirement Age." *Industrial and Labor Relations Review*, January 1980, 33(2), pp. 241-55.
- Honig, Marjorie. "The Subjective Probabilities of Retirement of White, Black, and Hispanic Married Women." HRS Working Paper 94-008, Institute for Social Research, University of Michigan, Ann Arbor, 1994.
- _____. "Retirement Expectations Over Time." HRS/AHEAD Working Paper 96-038, Institute for Social Research, University of Michigan, Ann Arbor, 1996.
- Hurd, Michael D. "The Joint Retirement Decision of Husbands and Wives," in David A. Wise, ed., *Issues in the economics of aging*. Chicago: University of Chicago Press, 1990, pp. 231-54.
- Hurd, Michael D. and McGarry, Kathleen. "Evaluation of Subjective Probability Distributions." HRS Working Paper 94-004, Institute for Social Research, University of Michigan, 1994.
- Pozzebon, Sylvia and Mitchell, Olivia S. "Married Women's Retirement Behavior." *Journal of Population Economics*, January 1989, 2(1), pp. 39-53.
- Shaw, Lois B. "Retirement Plans of Middle-Age Married Women." *Gerontologist*, February 1984, 24(2), pp. 154-59.

Gender Differences in the Allocation of Assets in Retirement Savings Plans

By ANNIKA E. SUNDÉN AND BRIAN J. SURETTE*

In 1995, 40 percent of working men and 32 percent of working women were covered by a defined contribution (DC) plan. A distinguishing characteristic of these plans is that workers can generally choose how their assets are invested. Using data from the 1992 and 1995 Surveys of Consumer Finances (SCF), this paper examines whether workers differ systematically by gender in the allocation of assets in DC plans. Previous researchers have reported that many workers tend to invest their retirement assets too conservatively, and in particular that women are less likely than men to invest in risky assets such as stocks. In the presence of an equity premium, a lower propensity by women to invest in stocks could translate into large differences in the accumulation of financial wealth for retirement. We establish that gender differences in investment decisions exist, though they are more complicated than previous studies have suggested. We show that these differences are not completely explained by differences in individual or household characteristics.

A few studies have examined gender differences in investment decisions (Vickie L. Bajtelsmit and Jack L. VanDerhei, 1997; Richard P. Hinz et al., 1997). These studies use administrative data and report that women tend to invest their retirement funds in less risky vehicles than men. Michael Haliassos and Carol C. Bertaut (1995) use the 1983 SCF to examine why such a large fraction of households do not own any stock. They report that gender does not have a significant effect on the probability of owning stock, though gender differences are not the focus of their paper.

What these data sources lack (Haliassos and Bertaut being the exception) is a rich set of demographic and other variables on households that theory predicts should affect investment behavior. This paper adds to the literature by examining gender differences in investment decisions conditioning on such variables. The results highlight the importance of including marital status, risk-aversion measures, and the portfolio of assets held outside DC plans when examining gender differences in investment decisions in these plans.

I. Data

The data used in this paper come from the 1992 and 1995 Surveys of Consumer Finances, a triennial survey sponsored by the Federal Reserve Board in cooperation with Statistics of Income. The SCF collects detailed information on households' assets, liabilities, and demographic characteristics as well as on pension coverage, pension plan characteristics, and the allocation of assets in DC plans.¹ The survey sample size was 3,906 households in 1992 and 4,299 households in 1995. Descriptive statistics on pension coverage are presented in Table 1.

These data enable us to undertake a detailed analysis of investment choices in DC plans and relate it to individual and household characteristics. Most information is collected at the household level. However, data on pension coverage, employment, and other demographic characteristics are available for both the household head and the spouse/partner. We use person-specific information to split married households into two observations. Variables collected at the household level, such as financial wealth, are attributed to both

* Board of Governors of the Federal Reserve System, 20th/C Street, N.W., Stop 153, Washington, DC 20551. We thank Raphael Bostic, Bruce Fallick, Arthur Kennickell, and Martha Starr-McCluer for helpful comments. The views presented are those of the authors alone and do not necessarily reflect the official position of the Board of Governors.

¹ Arthur B. Kennickell et al. (1997) describe the SCF in detail.

TABLE 1—PENSION COVERAGE IN 1992 AND 1995

Statistic	1992		1995	
	Men	Women	Men	Women
Percentage with pension	56.8	50.2	56.5	48.0
Percentage with DB plan	34.9	29.4	24.8	21.8
Percentage with DC plan	31.9	28.0	40.1	32.2
Percentage with DC plan who can direct investment	NA	NA	71.4	69.8
Investment in DC plan				
Mostly stock	32.2	28.6	40.2	39.1
Mostly bond	27.9	29.5	16.1	21.2
Diversified	40.0	41.9	43.7	39.8
Median amount in DC account (\$1995, thousands)	10.8	5.4	10.4	5.5
Number of observations	1,443	1,380	1,705	1,669

Notes: DB = defined benefit; DC = defined contribution. Sample: Working men and women, age less than 75 (Survey of Consumer Finances, 1992 and 1995).

of these person records while person-specific information is attributed to the individuals separately.² We believe this is reasonable since married couples can draw on shared finances. The sample consists of individuals currently working, covered by a DC plan and under the age of 75.³

II. Model

The investment choices reported in the SCF for defined contribution plans are categorical: (i) invest mostly in stocks; (ii) invest mostly in interest earning assets (hereafter “bonds”); and (iii) investments split between stocks and

interest-earning assets (hereafter “diversified”). The SCF does not collect information on the specific allocation of plan assets, and we cannot derive portfolio shares. Since no clear ordering exists among the alternatives, a multinomial logit model is used to analyze investment behavior.

To estimate the model we pool data from the 1992 and 1995 SCF's.⁴ The first column of Table 2 contains descriptive statistics for the pooled sample. We assume that each individual ultimately decides how his or her retirement assets are invested. To account for the possibility that married individuals may coordinate their investment decisions, and that the effects of gender may differ by marital status, we include indicator variables for gender and marital status, as well as an interaction variable of the two.

Other research has argued that financial knowledge is an important determinant of investment decisions (B. Douglas Bernheim and Daniel M. Garrett, 1996). To proxy for financial knowledge we include indicator variables for levels of schooling.

The allocation of assets within (and outside) retirement plans is likely to be correlated with the willingness of households to trade risk for return. We control for attitudes toward risk by including self-reported measures of each household's willingness to exchange risk for return.⁵

Savings in defined-contribution plans are only one part of households' portfolios. The overall level of financial risk and return facing a household depends on the mix of all its financial assets. Decisions about how to invest defined-contribution savings should, therefore, depend on the financial assets held outside DC accounts. To address such factors, we include variables to control for the share of each household's other savings that are held in stocks, bonds, and in other financial assets.

² This data strategy may result in nonindependent observations and may cause regression estimates to be inefficient.

³ The SCF first collected data on self-direction of DC assets in 1995. These data show that those who can and cannot direct their investments do not differ in their allocation decisions. We therefore use all DC participants in 1992 and 1995.

⁴ The results are qualitatively similar (though less precise) when the model is estimated for the two years separately.

⁵ This question is asked only for the household as a whole; we think it likely that members of the same household have similar risk-return preferences.

These shares also include assets held in Individual Retirement Accounts (IRA's).⁶

Several of the variables included in the model are likely to be endogenous or simultaneously determined with DC investment decisions. These econometric issues are very difficult to address. The analysis below does not account for these complications. Thus, we caution the reader that the results are descriptive rather than causal.

III. Results

The results of our analysis are presented in Table 2. We chose the "diversified" category as the base category for the multinomial logit. Therefore, the effects of each variable described below refer to its effect on the probability of choosing "mostly stocks" or "mostly bonds" relative to the probability of choosing the "diversified" category. To simplify the exposition, we discuss the results without reference to this normalization.

The results demonstrate that it is not gender alone that determines investment choice. Rather, investment decisions seem to be driven more by a combination of gender and marital status. According to the estimates, single women and married men are less likely than single men (the comparison group) to choose "mostly stocks." Though the interaction of marital status and gender is statistically significant, a joint test of all three gender-marital coefficients indicates that married women do not differ significantly from other groups in their probability of choosing "mostly stocks." The estimates also indicate that, though women and men do not differ, married women are more likely than single women to choose "mostly bonds."

These results demonstrate that the effects of gender on investment decisions are more complicated than previous research has suggested. In fact, ignoring the possibility that the effects of gender differ by marital status would lead to very different conclusions. Estimates from models that omit the gender \times marital interaction term (not reported) suggest that marital

TABLE 2—DESCRIPTIVE STATISTICS, MULTINOMIAL LOGIT, AND PROBIT RESULTS (STANDARD ERRORS IN PARENTHESES)

Characteristic	Mean	Mostly stocks ^a	Mostly bonds ^a	Have a DC ^b
Demographics				
Female	0.433	-0.567* (0.251)	-0.369 (0.322)	0.148† (0.083)
Married	0.817	-0.491* (0.227)	-0.291 (0.283)	0.095 (0.073)
Female \times married	0.322	0.740* (0.278)	0.791* (0.381)	-0.349* (0.091)
Age	42.28	-0.002 (0.040)	-0.003 (0.050)	0.069* (0.011)
No high-school degree	0.043	0.177 (0.311)	0.568† (0.337)	-0.223* (0.080)
Some college	0.231	-0.103 (0.167)	-0.039 (0.195)	0.047 (0.050)
College graduate	0.485	0.130 (0.155)	0.096 (0.183)	0.145* (0.054)
Risk/return preferences				
Above average	0.283	0.407* (0.175)	-0.547* (0.213)	
Average	0.487	0.077 (0.156)	-0.418* (0.164)	
Portfolio share of nonretirement assets^c				
Bond Holdings				
0-33 percent	0.844	-0.994† (0.577)	-0.417 (0.691)	
33-67 percent	0.108	1.752† (0.995)	-0.465 (1.36)	
67-100 percent	0.048	1.291 (2.06)	4.273 (2.70)	
Stock holdings				
0-20 percent	0.648	1.185 (0.99)	-0.930 (1.25)	
20-80 percent	0.288	-0.043 (0.483)	-1.362* (0.650)	
80-100 percent	0.063	2.331 (2.53)	6.424† (3.31)	
Number of observations	2,098	2,098	2,098	6,197

Note: Standard errors account for imputation variance.

^a Also includes age-squared, tenure, income, financial, nonfinancial, and IRA assets, total debt, number of children under age 12, percentage invested in "other" assets, and dummies for having no financial assets, year, race, homeownership, and whether household has rights to a DB plan.

^b Also includes the variables listed in footnote ^a plus occupation and full-time-status dummies.

^c Instead of means, we report the percentage of individuals in each portfolio share category.

† Statistically significant at the 10-percent level.

* Statistically significant at the 5-percent level.

status has no effect on investment decisions, and that women have a higher probability of choosing "mostly bonds." The results in

⁶ Only the household's overall allocation of IRA's is collected.

Table 2 show clearly that marital status matters and that it interacts in important ways with gender. Surprisingly, neither education nor age seems to affect allocation decisions.

The risk-preference measures have the expected effects on portfolio choices. A willingness to take above-average risk for above-average return increases the probability of choosing the "mostly stocks" category and reduces the probability of choosing the "mostly bonds" category. Workers willing to take average risk in exchange for average returns are also less likely to choose "mostly bonds."

To account for portfolio effects, we include a kinked, linear spline function of the percentages of financial assets held in stocks and bonds.⁷ Table 2 shows that individuals with less than 33 percent of their financial assets in bonds are less likely to choose "mostly stocks." The estimates also show that the probability of holding "mostly bonds" declines as the percentage of financial assets held in stocks rises from 20 percent to 80 percent. This effect reverses dramatically once individuals reach stock allocations of 80 percent.

These portfolio effects are largely consistent with our expectations. However, there is also evidence of an unexpectedly persistent preference for bonds: individuals with large allocations of financial assets to bonds are more likely to invest their DC assets in "mostly bonds." Given the large number of covariates included in these regressions to explain life-cycle factors and risk preference, this result suggests that some individuals are highly averse to investing more than a small percentage of their financial assets in stocks. It is telling that individuals who choose mostly bonds are much more likely to state that they are unwilling to take any financial risk, even though investing exclusively in bonds exposes the household to considerably more real-interest-rate risk.

The gender and marital differences in investment behavior described above may par-

tially be driven by self-selection into jobs with DC plans. It is not clear how a model addressing this econometric problem would be identified. For this reason, we did not control for selection in the allocation model. However, we think it is important to discuss participation in DC plans.

The data show that women are less likely than men to have DC plans (Table 1), and that married women are least likely to have such plans (not shown). If these differences persist, women may end up accumulating less wealth for retirement regardless of how they invest their DC assets. To test whether gender and marital differences in DC participation remain in a multivariate framework, we estimate a probit of DC participation on the pooled 1992–1995 SCF sample. The results are reported in the last column of Table 2. The results indicate that single women are more likely than single men to have a DC plan. Married women, however, are much less likely than men (or single women) to have a DC plan.

The salient characteristics of jobs that offer DC plans are difficult to identify precisely. We include occupation and full-time dummies in the participation probit to proxy for job characteristics that might explain gender and marital differences in coverage. Although not reported in the table, both measures significantly affect DC participation. Professional and "skilled white-collar" workers are about 5-percent more likely than "unskilled blue-collar" workers to have a DC plan, while service workers are about 11-percent less likely than "unskilled blue-collar" workers to be covered. Full-time workers are about 26-percent more likely than part-time workers to have a DC plan. We hypothesize that measurement difficulties coupled with gender and marital differences in occupational choice and full-time status may partially explain why women's DC participation has increased at a slower rate than men's and why gender and marital differences in participation persist in a multivariate framework.

IV. Conclusions

We conclude that gender and marital status significantly affect how individuals choose to al-

⁷ The kink points were selected to minimize trends in predicted probability "residual" plots for small groupings of observations sorted by percentage allocations to stocks or bonds. The qualitative results are not overly sensitive to the selection of kink points.

locate assets in defined-contribution plans. We control for a wide range of demographic, financial, and attitudinal characteristics that previous researchers have argued could explain such differences. Our results indicate that such controls are important but do not explain away gender and marital effects. Because these controls are imperfect, however, and because unobserved differences may affect investment behavior, we interpret the remaining gender and marital differences with caution. We view them as descriptive, rather than causal.

The trend toward defined-contribution plans makes individual investment decisions particularly important in determining how much wealth is accumulated for retirement. In the presence of an equity premium, the failure of some groups, such as single women, to invest sufficient assets in stocks may lead to lower retirement wealth. Moreover, some of the proposed reforms of the Social Security system will allow workers to choose how their Social Security contributions are invested. Our results, therefore, shed some light on how public and private retirement wealth may be distributed in the future.

REFERENCES

- Bajtelsmit, Vickie L. and VanDerhei, Jack L. "Risk Aversion and Pension Investment Choices," in Michael S. Gordon, Olivia S. Mitchell, and Marc M. Twinney, eds., *Positioning pensions for the twenty-first century*. Philadelphia: University of Pennsylvania Press, 1997, pp. 45-66.
- Bernheim, B. Douglas and Garrett, Daniel M. "The Determinants and Consequences of Financial Education in the Workplace: Evidence from a Survey of Households." National Bureau of Economic Research (Cambridge, MA) Working Paper No. 5667, July 1996.
- Haliassos, Michael and Bertaut, Carol C. "Why Do So Few Hold Stocks?" *Economic Journal*, September 1995, 105(432), pp. 1110-29.
- Hinz, Richard P.; McCarthy, David D. and Turner, John A. "Are Women Conservative Investors? Gender Differences in Participant Directed Pension Investments," in Michael S. Gordon, Olivia S. Mitchell, and Marc M. Twinney, eds., *Positioning pensions for the twenty-first century*. Philadelphia: University of Pennsylvania Press, 1997, pp. 91-103.
- Kennickell, Arthur B.; Starr-McCluer, Martha and Sundén, Annika E. "Family Finances in the U.S.: Recent Evidence from the Survey of Consumer Finances." *Federal Reserve Bulletin*, January 1997, 83(1), pp. 1-24.

How Are Participants Investing Their Accounts in Participant-Directed Individual Account Pension Plans?

By LESLIE E. PAPKE*

The fastest growing type of defined contribution (DC) plan is the 401(k)-type plan. These plans have a cash or deferred arrangement (CODA) feature that allows participants to make pretax contributions. Assets in 401(k) plans exceed \$440 billion for 19.1 million participants (U.S. Department of Labor, 1995). Participants in these plans typically choose their own investments, but non-CODA defined contribution plans may permit employees to select their own investments as well. Thus, the asset allocation in a defined contribution plan plays a critical role in the adequacy of retirement income.

While sponsor investment choices for defined benefit plans have been studied extensively (see Papke [1992] for a review), only recently has research turned to the asset choices of participants in defined contribution plans. The issue has been prevalent in the popular press, however, as reporters have sounded alarms about "excessively conservative" investments by participants in 401(k) plans, especially by female participants (Mary Rowland, 1995), and general participant insecurity about investment choices. Findings from recent econometric studies differ with regard to systematic gender differences in asset choices (see James M. Poterba and David Wise, 1996; Vickie L. Bajtelsmit and Jack L. VanDerhei, 1997; Richard P. Hinz et al., 1997). Since conservative investments historically earn lower rates of return than riskier assets, a persistent gender difference in asset choices could create pension benefit differences even between similarly situated participants. But there has been no research to date that examines the effect of asset choice as well as gender on investment patterns. This paper

begins to fill that gap with evidence from the 1992 National Longitudinal Survey of Mature Women.

I. Gender and Asset Choice: The NLS Mature-Women Survey

To estimate the effect of gender and participant direction on DC plan assets, I use the 1992 National Longitudinal Survey (NLS) of Mature Women, an ongoing survey that began in 1967 with 5,083 women ages 30–44 (3,094 were included in the latest survey). In addition to detailed demographic data on the respondent and her household, supplementary questions on her and her spouse's pension eligibility and benefits from current/past employers or from other pension sources (e.g., personal plans) were included in the 1979, 1982, 1986, and 1989 questionnaires. The 1992 questionnaire expanded the set of pension questions to include multiple pension coverage from (i) future pensions from current employers; (ii) current pensions from previous employers; and (iii) future pensions from previous employers.

Information on DC plans includes the type(s) of plan (e.g., thrift/savings, 401(k), profit-sharing, stock purchase, or other), dollar amounts that both employer and respondent contributed, and the employee's contribution percentage, account balance, and how the dollars were invested. Specifically, respondents were asked: Were you able to choose how the money in your account is invested? How is the money in this account invested? Is it mostly in stocks, mostly in interest-earning assets, is it split evenly between these, or what? The three responses are mostly (51 percent) or all stocks; mostly (51 percent) or all interest-earning assets; and split evenly between the two (see Papke [1997] for a detailed description).

I restrict the sample to participants in DC plans. There are 232 participants in such plans

* Department of Economics, 101 Marshall Hall, Michigan State University, East Lansing, MI 48824. I thank Denise Takahashi Heidt for excellent research assistance.

and 204 separate families. That is, only 28 of the plans have duplicate identifiers indicating a second plan for the family—either a second plan for the respondent, or two plans for the spouse, or one for each spouse. The sample is further restricted to those who respond to the investment questions. Women's plans comprise 59 percent of the sample, and the plans of married women comprise 33 percent of the sample. Choice is available for 62 percent of the plans, and 21 percent of plans have a profit-sharing component (i.e., the employer's contribution is determined in part by profits). The average account balance is \$27,830, and the average employee contribution percentage is 4.68 percent of salary.

Because the survey is of mature women, the sample is not representative of the pension-age population in the United States. In particular, the sample is older and excludes single men, so that the findings may not generalize to younger people. (The two NLS surveys that focused on males [the National Longitudinal Survey of Older Men, administered from 1966 to 1983, and the Survey of Young Men, administered from 1966 to 1981] do not include many pension questions and include no investment information.) Still, this is an interesting group to look at since these people, nearing the end of their professional careers, are likely to have pensions and to consider retirement income seriously.

II. Econometric Results for Percentage Investment in Stocks

This section presents econometric evidence on how choice and demographic variables affect asset allocation in DC plans. Recall that the response categories are mostly bonds, mixed, and mostly stocks. This is a discrete ordered outcome, and so ordered response methods (such as ordered logit and ordered probit) are appropriate. Estimating an ordered-response model allows estimation of the effect of choice and other variables on the probability of each response but not of the effect on the percentage invested in stock. The estimated coefficients from ordered logit or probit models give directions of effects, but even these magnitudes are difficult to interpret. This can be overcome by calculating partial effects

at various values of the explanatory variables, but the information is difficult to summarize.

Because linear models are easy to interpret and are likely to give good estimates of the average effects, I focus on linear regression. This requires defining a "percentage invested in stocks" variable. I follow the convention used in the existing literature and code the response "mostly stocks" as 100 percent, the response "mixed" as 50 percent, and the response "mostly bonds" as 0 percent invested in stocks.

Choosing the extreme values of 0 and 100 likely overestimates the effect of any variable on the true percentage invested in stock. While in simple cases one can compute the bias (see Papke, 1997), in general, one cannot know by how much the effect will be overstated. While this results in a mismeasurement of the true percentage invested in stocks, it provides a useful base case.

Given the estimates for this base case, it is easy to obtain the estimates for a symmetric reassignment of the extreme values. For example, if 25, 50, and 75 are chosen as the numeric responses, the slope coefficients I report are divided by two; the *t* statistics do not change. (The general formula for rescaling the reported slope coefficient for a symmetric reassignment c , 50 , $100 - c$ for c between 0 and 50 is $1 - [c/50]$.) Statistical significance does not hinge on the definition of percentage in stock. To be sure that the conclusions are not influenced by using a linear model, I also estimate ordered logit models and compare the signs and significance of the coefficients.

Table 1 presents ordinary least-squares (OLS) estimates of the percentage invested in stock. Because the dependent variable is discrete, I report heteroscedasticity-robust standard errors. In addition, about 14 percent of the identifiers have two observations. Therefore, the standard errors are corrected for within-family (identifier) correlation. (Random-effects estimation produces very similar results to OLS. Because of the small number of duplicate observations and the lack of variation across family in the explanatory variables, fixed-effects estimation is not informative.)

The basic regression includes demographic and plan characteristics, but no financial

TABLE 1—PARTICIPANT-LEVEL MODELS FOR THE PERCENTAGE OF DEFINED-CONTRIBUTION ASSETS IN EQUITIES

Independent variable	Regression		
	(i)	(ii)	(iii)
Constant	121.11 (49.47)	115.66 (53.91)	95.67 (54.18)
Female ^a	0.39 (7.88)	0.34 (8.80)	3.02 (8.92)
Female × married	0.13 (7.41)	2.09 (8.33)	2.10 (8.42)
Choice	14.34 (6.00)	14.77 (6.46)	14.91 (6.85)
Age	-1.57 (0.72)	-1.55 (0.79)	-1.20 (0.79)
Education	0.67 (1.01)	0.73 (1.11)	.38 (1.14)
Profit-sharing	25.92 (11.12)	23.38 (11.27)	25.08 (11.65)
Profit-sharing × choice	-19.75 (14.81)	-15.71 (15.15)	-19.86 (15.42)
Income 25–50K		8.87 (8.25)	13.65 (7.96)
Income 50–100K		2.20 (10.86)	6.90 (10.59)
Income >100K		-14.14 (11.03)	-8.91 (11.58)
Net worth 50–100K		-7.45 (9.49)	-10.58 (9.67)
Net worth 100–250K		3.10 (8.27)	-3.20 (8.49)
Net worth 250–500K		3.21 (10.52)	-4.55 (10.83)
Net worth >500K		-4.30 (12.21)	-11.97 (13.21)
Spouse holds equities in DC plan (percent)			0.16 (0.15)
Stocks in 1989 (\$1,000's)			0.31 (0.16)
Number of observations:	201	194	182
Adjusted R ² :	0.042	0.031	0.048

Notes: Standard errors that are robust to heteroscedasticity and to correlation within family are in parentheses.

^a Only men who are spouses of women in the 1992 NLS Mature Women Survey are included in the sample.

information. There is no evidence of gender differences in investment in stock either for single or for married women (recall, all men in the sample are married). The effect of choice is large and statistically significant, indicating that people who are able to choose their investments put 14 percentage points more in stock than people without choice (assuming the plan is not profit-sharing). This assumes that the effect of choice is the same for men and women. When I allow for gender differences, the effect of choice for men is 17.71 with a standard error of 9.21, and the effect for

women is 12.58 with a standard error of 7.33. The difference, -5.13, is not statistically significant; its *t* statistic is -0.45. Subsequent regressions assume that the effect of choice is the same for men and women.

I find (as others have) some evidence that investment in stocks declines with age: the coefficient estimate indicates about a 2-percentage-point drop in stock for each additional year of age. This is evident despite the fact that the sample is older (75 percent of the sample is between the ages of 55 and 75). An extra year of education has no statistically significant effect on investment patterns, although the coefficient is positive.

DC plans are not a homogenous group. Of particular relevance for a regression of investment in stocks, profit-sharing plans often distribute the employer's contribution in shares of stock. I include a profit-sharing dummy as well as one interacted with choice. As expected, profit-sharing plans are predicted to allocate almost 26 percentage points more in stock if the participants have no choice. But the effect of choice in profit-sharing plans is negative; participants put over 5.41 percentage points less into stocks. This difference is not statistically significant; people put much more in stock in a profit-sharing plan, and having choice is irrelevant.

The second specification in Table 1 adds categorical variables for family income and net worth. No consistent pattern emerges, and no financial coefficient is statistically significant on its own (an *F* test of all seven variables has a *p* value of 0.16). There is a peculiarly negative coefficient for income above \$100,000; this is the one income category where previous researchers found a positive effect of income. The effect of choice is unchanged.

The final column of Table 1 adds two variables designed to capture the potential for diversification within the family and a family taste for stocks. The first variable added is the percentage of stock that one's spouse holds in his/her DC plan. The coefficient is 0.16, indicating a very small effect that runs counter to diversification within families and is not statistically significant at conventional levels. The second variable is the amount of stock the family held outside of pension plans in 1989

(the year the National Longitudinal Survey of Mature Women included detailed financial questions). The coefficient of 0.31 does indicate a family taste for stocks; for every \$1,000 of stock held in 1989, the participant holds almost one-third of a percentage point more in stock in his/her DC plan.

I also estimate ordered logit equations (not reported here). They are consistent with the linear-regression results. The statistically significant coefficients have the same sign and are still statistically significant.

III. Conclusions and Caveats

Contrary to recent popular wisdom, evidence presented here suggests that participants choose to invest more in stocks when given the choice. For the older group of participants I study here, there appear to be no differences in investment patterns by gender.

One might argue that choice is endogenous in these regressions. That is, participants inclined to risky investment join firms that offer plans with choice. By including many demographic covariates, I have tried to control for characteristics that incline the participant toward stocks. This potential problem is pursued in more detail in Papke (1997). I find no partial correlation between the knowledge of the pension plan (an additional survey question) and investment in stocks, suggesting that more informed participants are not necessarily inclined toward stock investment.

Further, there are several data limitations. First, the asset categories in the National Longitudinal Survey of Mature Women are not detailed. One cannot determine with these data, for example, whether the equity investment is in employer stock, which would have different implications for the participant's personal investment risk than an investment in a diversified stock fund, since his or her employment fortunes and retirement income are tied to the same employer. This shortcoming is common to available data sets.

Second, it is possible that, even with choice, a substantial fraction of the contribution is constrained. Often the employer contribution is made in company stock, while the participant would still report choice over her or his

investment. Then, a larger fraction of the pension might be in equities than the participant would choose, unconstrained.

Third, the sample does not include single men, so I am unable to examine more closely the interaction between gender and marital status which may complicate investment choices. Finally, a female participant in this survey may have answered pension questions about her husband's plans without the presence of her spouse. At the time of the 1992 survey, the NLS did not record whether the spouse was actually at the interview, or whether the interviewer exercised the option of contacting the husband directly. It is not possible to determine what effect, if any, this may have had on the responses.

REFERENCES

- Bajtelsmit, Vickie L. and VanDerhei, Jack L. "Risk Aversion and Pension Investment Choices," in Michael S. Gordon, Olivia S. Mitchell, and M. Marc M. Twinney, eds., *Positioning pensions for the twenty-first century*. Philadelphia: University of Pennsylvania Press, 1997, pp. 45-66.
- Hinz, Richard P.; McCarthy, David D. and Turner, John A. "Are Women Conservative Investors? Gender Differences in Participant-Directed Pension Investments," in Michael S. Gordon, Olivia S. Mitchell, and Marc M. Twinney, eds., *Positioning pensions for the twenty-first century*. Philadelphia: University of Pennsylvania Press, 1997, pp. 91-103.
- Papke, Leslie E. "The Asset Allocation of Private Pension Plans," in John A. Turner and Daniel Beller, eds., *Trends in pensions 1992*. Washington, DC: U.S. Department of Labor, Pension and Welfare Benefits Administration, 1992, pp. 449-81.
- _____. "Asset Allocation in Participant-Directed Individual Account Pension Plans." Mimeo, Michigan State University, 1997.
- Poterba, James M. and Wise, David, D. "Individual Financial Decisions in Retirement Saving Plans and the Provision of Resources for Retirement." National Bureau of Economic

Research (Cambridge, MA) Working Paper No. 5762, September 1996.

Rowland, Mary. "Taking the Power of the 401(k), and Handing it to Someone Else." *New York Times*, 18 June 1995, section 3, p. 5.

U.S. Department of Labor. "Private Pension Plan Bulletin: Abstract of 1991 Form 5500 Annual Reports." Bulletin No. 4, Pension and Welfare Benefits Administration, U.S. Department of Labor, Washington, DC, Winter 1995.

LIFE-CYCLE AND COHORT STUDIES OF AGING[†]

Aging in the Early 20th Century

By CLAYNE L. POPE AND LARRY T. WIMMER*

There is a new awareness of an evolution in human "physical" capital that may rival changes in human "intellectual" capital and changes in technology in its impact upon economic growth and ultimately upon our standard of living or quality of life (Robert W. Fogel, 1994). Of particular interest are the dramatic improvements in morbidity and mortality over the past 100 years, especially that part of this improvement that began well before major investments in public health or gains in modern medical technologies. These documented improvements in human "physical" capital seem to be both cause and effect of the earlier Industrial Revolution, which appears to have produced its own revolution in nutrition, health, and longevity.

At the same time, there are many new questions raised by these observed improvements, including the nature of the relationship between various anthropomorphic measures and the risk of disease and death; the pace and timing of the observed cycling and eventual decline in height and rise of mortality during the mid and late 19th century, a period of rising real wages and increasing per capita income; and the impact of these changes upon one's ability to leave the labor force for leisure or retirement.

Cross-sectional data sets are unlikely to provide good evidence of past changes or future developments in morbidity and mortality. Life expectancy at birth based on a cross section of mortality rates for a given year has proved to be a poor estimate of actual mortality experience. Both a well-founded understanding of long-term trends in human health and a solid

empirical basis on which to make policy decisions about the future require development and analysis of longitudinal data sets. Furthermore, such data sets must be very long-term in nature in order to establish possible relationships between conditions and events in childhood and late-age experience and behavior. Biological processes imply that cohort effects on morbidity, mortality, and related issues such as health expenditures and retirement are likely to be significant and that the effects could be delayed. Consequently, it may well be necessary to examine variables and conditions that are more than 75 years apart.

One of the most promising new data sets which may shed considerable light on these and other questions related to changes in human "physical" capital is the Union Army recruit sample. This sample consists of approximately 36,000 men born between 1820 and 1843, and it forms the core of the NIH-NSF-funded "Early Indicators of Later Work Levels, Disease and Death" project. For the past six years the Department of Economics at Brigham Young University and the Center for Population Economics at the University of Chicago have been involved in the collection, cleaning, and coding of records from seven different data sources which provide impressive life profiles of these recruits and their families. The cornerstone of this new data set is the pension system of the Union Army. These pension records (PR) contain extensive information on occupational patterns, geographic mobility, health (including diseases), accidents and injuries, family structure, living arrangements and conditions after retirement, and cause of death. To date, information on 303 randomly selected companies of the Union Army has been collected, covering over 36,000 Civil War recruits, more than 26,000 of whom are found in the pension system, and almost 20,000 with periodic

[†] Discussants: David Weir, University of Chicago; Joel Mokyr, Northwestern University.

* Department of Economics, Brigham Young University, Provo, UT 84602, and NBER.

medical examinations. In addition to these pension records, the data set contains the military service records for all recruits, wartime medical records of those entering Army hospitals, and data from the Census manuscripts for 1850, 1860, 1900, and 1910. Thus, in addition to an almost continuous record of employment and migration, the final data set will include frequent medical examinations by three board-certified physicians. These examinations of general health conditions, as well as complaints listed by the recruit, occur at an average interval of three years once a pensioner enters the system. Ratings given by the board physicians represent the severity of observed health problems. The pension records and the Census manuscripts are used to record the family structure, including the ethnicity, occupation, and wealth of the recruit's parents.

We are currently in the final year of collection of this data set which will ultimately be available for public use in several forms, including CD ROM with accompanying documentation. Just how extensive the coverage may be for a typical recruit is indicated by the "potential" number of variables within each record: military service record (MSR), 799 variables; carded military record (CMR), 795 variables; pension record (PR), 1,352 variables; Census records (CEN), 1,294 variables; and the medical examinations (Surgeons Certificates [SCRT]), 2,007 variables, for a total number of 6,247 possible observations per recruit. For cohorts of white males born between 1840 and 1844, between 75.1 percent and 97.7 percent were examined for military service, and between 57.8 percent and 81.4 percent actually served. Findings from earlier work verify that these Union Army recruits are representative of the Northern, white-male population of military age; that the foreign-born served in approximately the same proportions as native-born; that recruits came from households with the same average wealth as the Northern male population as a whole; and that this sample reflects the geographic distribution of the northern population (Fogel and Wimmer, 1992).

Proof of the value of this new data set for longitudinal studies of factors affecting the aging process, rests on the quality and significance of the work produced. This paper

examines the work of five scholars who have been among those principally involved over the past six years, not only in the construction of the data set, but in using the Union Army data to examine an interesting range of problems associated with aging. Four of these scholars, Chulhee Lee, Dora L. Costa, John Kim, and Sven Wilson, have been graduate students of Robert Fogel, the principal investigator of the "Early Indicators . . ." project, and one, Louis Nguyen, was at the time a medical student at the University of Chicago. Much of the important work to date stemming from the Union Army sample may be summarized by the previous work of these five scholars.

The basic strength of the Union Army sample lies in the linkage of anthropomorphic measures to mortality and morbidity. John Kim, in his recent dissertation (Kim, 1996), examines the regularities observed in the weight-risk and height-risk relationships among different populations. Building upon the earlier work of Hans Th. Waaler (1984), Kim adds evidence of the value of height-weight-risk surfaces and the body-mass index (BMI) in tracking changes in health and mortality trends. Kim's findings of the independent relationship of both height and BMI upon mortality, and of the secular trend in the average level of BMI have important policy implications. Given our inability to change height beyond childhood, it requires that we examine policies regarding prenatal and early childhood care and nutrition if a nation is to move toward a more optimal BMI.

Dora Costa's (1995a, b, 1996) work on factors affecting the age of retirement has added considerably to our knowledge of that process. She finds strong evidence that the tendency toward retirement and earlier ages of retirement began well before public-policy legislation of the New Deal. Between 1880 and 1930 labor-force participation of men 65 years of age and older dropped from 78 percent to 58 percent (Jon Roger Moen, 1987). By comparing retirement rates found in the Civil War pension records with current rates, Costa found that 70 percent of the decline in labor-force participation experienced among white males between 1880 and 1990 occurred before 1960, and much of this occurred before the passage

of the Social Security Act (Costa, 1995a p. 298). Clearly, the Civil War pension played a role in the decision of its recipients to retire. By 1900, the Civil War pension was received by 35 percent of all U.S. white males between the ages of 55 and 59 and 21 percent of those aged 60–64. Pension income averaged \$135 per year or 53 percent of the income of farm laborers and 36 percent of the wage of non-farm workers (Costa, 1995a p. 299). Even though the Civil War pension system did not require that one leave the labor force, as in the case of modern Social Security benefits, Costa found that a \$10 increase in monthly pension income increased the probability of retirement by 0.09. The retirement rate among Civil War veterans in the pension system was double that of a random sample of nonveterans. Perhaps the most important finding for modern policymakers is the declining importance of income over time on the decision to leave the labor force. The elasticity of labor-force participation with respect to pension income was very strong in the Union Army sample, at 0.73, and rising as income from pensions rose. Other recent studies find a smaller impact on labor-force nonparticipation whether compared against assets or Social Security retirement and disability payments, with measured elasticities varying between 0.16 and 0.63 (Costa, 1995a pp. 315–16). These findings suggest the following conclusions: first, at the turn of the century there existed a strong relationship between income and retirement; second, the importance of current income or wage upon the retirement decision has been declining over time; third, the importance of leisure goods upon retirement decisions appears to be increasing; and finally, the Social Security administrators and lawmakers may face considerable difficulty in attempting to induce later ages of retirement as a remedy for the financial problems of Social Security.

Costa's work on height, weight, and mortality, suggests the value of the Union Army recruit sample in the modern debate involving anthropomorphic issues (Costa, 1993a, b). Costa finds in the Civil War sample an average BMI of 22.8 compared to 25.0 for males in Waaler's study of modern Norway (Costa, 1993a p. 441). This low BMI among pensioners of the Civil War partially explains why

their mortality rate was higher than that of cohorts born today. Men with a BMI below 20 made up 14 percent of the sample and accounted for 38 percent of total mortality. If Union Army recruits had possessed the modern BMI distribution found in Waaler's study, the resulting 14-percent decline in mortality would account for 20 percent of the total decline between 1900 and 1986 (Costa, 1993a p. 444).

Combining Costa's findings relating BMI and mortality with her earlier studies of income and retirement has produced important results (Costa, 1996). Not only would a modern BMI have reduced mortality, but it would also have increased labor-force participation (by reducing retirement) by as much as 10 percent and, therefore, increased total output. Costa finds a strong relationship between retirement and health in the Union Army data. At the same time, however, in addition to the declining importance of current income, there exists even stronger evidence of a significant decline over time in the importance of ill health as an explanation for one's decision to leave the labor force. Costa finds the elasticity of labor-force nonparticipation with respect to ill health falling from 1.07 in the Civil War sample to 0.28 in the National Health Interview Survey (NHIS) sample (Costa, 1996 pp. 79–80).

Chulhee Lee's work with the wartime records of the Union Army recruits yield some equally surprising results (Lee, 1997). Previous studies examining general trends and conditions of health have identified the greater likelihood of early death for foreign-born compared to native-born individuals, of urban compared to rural dwellers, and nonfarm relative to farm inhabitants. The Civil War, however, created a situation in which many populations were brought together into an unusually dangerous disease environment, and according to Lee (1997), produced the opposite results from those observed for the general population. Native-born recruits from rural areas and farms suffered an increased risk of contracting disease and of dying from those diseases, compared to urban or foreign-born populations. Thus, those who were healthier before the war were at greater risk during the war, suggesting the importance of the general

disease environment upon health, and the mitigating effects of prior contact and natural immunities in the battle against disease. In addition, Lee also finds that prior family wealth does not provide explanatory evidence of war-related diseases and death among the Civil War recruits. The exceptions were those diseases that tend to be nutritionally based such as measles, diarrhea, and tuberculosis and other respiratory diseases. These results highlight the complex relationship between the disease environment and other socioeconomic characteristics.

Possibly one of the most important papers stems from an early 20-company subsample of the Union Army data. This paper (Fogel et al., 1993) finds evidence of higher prevalence rates for chronic disease among Civil War veterans compared to veterans of a similar age who served in World War II and the Korean Conflict. Heart disease was 2.9 times more prevalent in the Union Army subsample, musculoskeletal and respiratory diseases were 1.6 times as prevalent, and digestive diseases were 4.7 times greater than modern levels. These observed differences between comparable ages over the past 70 years suggest a decline in chronic disease rates as high as 6 percent per decade.

Work in progress by Sven Wilson, based upon a larger sample, suggests that further evidence will confirm the signs of the earlier results even though the magnitude of the decline is still unknown. If this intertemporal decline in chronic disease continues, it will have profound implications for improvements in the health and welfare of the U.S. population over time. These improvements in health and the increasing age of onset of chronic illness could reduce future demands for health care and increase the ability of an aging population to remain in the labor force. Based upon the work of Dora Costa, however, whether these improvements imply increased labor-force participation by those over 65 depends less upon their health than upon continued changes in income, demand for leisure, and the elasticity of nonparticipation with respect to income and health.

Wilson and Mark Rudberg (1996) are among the first to use the Union Army sample to examine disease-specific prevalence rates

by age. Their disease-specific results confirm the presence of a secular decline in chronic disease at older ages. They compare congestive heart failure (CHF) rates found in the physicians' examinations in the Union Army data with that found in the National Health and Nutrition Examination Survey (NHANES) I sample and in the Framingham Heart Study. From a "retrospective sample" of 752 Union Army pensioners alive on 1 January 1910, the incidence of CHF was 10.3 percent for individuals aged 65–74 and 22.2 percent for those aged 75–84. Analysis of the prevalence of CHF in the NHANES I sample for 1970 among those in the 65–74 age group resulted in a rate that was less than half that of the Union Army veterans, or 4.8 percent. Similarly, of 1,527 veterans followed by Wilson and Rudberg (1996) in an "1890 prospective sample," 161 developed CHF over a 20-year follow-up period. The resulting rate of 10.5 percent among the Union Army veterans compares to a rate of 7.6 percent found in the Framingham Heart Study over the period 1970–1990. In addition to the higher prevalence rates found among Union Army veterans, the work of Wilson and Rudberg provides important evidence of the usefulness of this new data set for the study of age-specific diseases and morbidity and mortality rates.

As useful as these findings are, we have only begun to explore the range of questions that may be addressed by this new data set. Among the issues remaining are the following: the effects of nutritional status, socioeconomic factors, and exposure to disease during developmental and middle age on the morbidity and mortality rates of white males at middle and late ages; the impact of exposure to warfare during late adolescence and early adulthood upon employment, morbidity, and mortality rates among white males who survive to middle and late ages; the effects of specific diseases and other disabilities on labor-force participation rates; the nature and cost-effectiveness of arrangements for care of the aged by the nature of their disabilities and the effects of the residence where they were lodged; the effects of health and wartime experience upon geographic and occupational mobility during middle and late ages; and differences between the cause of death implied

by an examination of cross-sectional death certificates and inferences from a longitudinal study of morbidity using the physicians' examinations. Data from this sample may also help resolve the nagging puzzle implied by the increase in life expectancy which began as early as 1870-1880, while height-by-age did not improve until as late as 1900. The above list is undoubtedly only a small part of the potential research agenda that may be addressed by the Union Army sample. This data set will help scholars shed new light on the beginnings of improvements in human "physical" capital as reflected by changes in health, mortality, and economic welfare during the late 19th and early 20th centuries.

REFERENCES

- Costa, Dora L. "Height, Weight, Wartime Stress, and Older Age Mortality: Evidence from the Union Army Records." *Explorations in Economic History*, October 1993a, 30(4), pp. 424-49.
- . "Height, Wealth, and Disease Among the Native Born in the Rural, Antebellum North." *Social Science History*, Fall 1993b, 17(3), pp. 355-84.
- . "Pensions and Retirement: Evidence from Union Army Veterans." *Quarterly Journal of Economics*, May 1995a, 110(2), pp. 297-320.
- . "Agricultural Decline and the Secular Trend in Retirement Rates." *Explorations in Economic History*, October 1995b, 32(4), pp. 540-52.
- . "Health and Labor Force Participation of Older Men, 1900-1991." *Journal of Economic History*, March 1996, 56(1), pp. 62-89.
- Fogel, Robert W. "Economic Growth, Population Theory, and Physiology: The Bearing of Long-Term Process on the Making of Economic Policy." *American Economic Review*, June 1994, 84(3), pp. 369-95.
- Fogel, Robert W.; Costa, Dora L. and Kim, John. "Secular Trends in the Distribution of Chronic Conditions and Disabilities at Young Adult Ages, 1860-1988: Some Preliminary Findings." Unpublished manuscript presented at the National Bureau of Economic Research (Cambridge, MA) Summer Institute, Economics of Aging Program, 26-28 July 1993.
- Fogel, Robert W. and Wimmer, Larry T. "Early Indicators of Later Work Levels, Disease, and Death." Working Paper Series on Historical Factors in Long Run Growth, National Bureau of Economic Research (Cambridge, MA) Historical Working Paper No. 38, June 1992.
- Kim, John. "The Economics of Nutrition, Body Build, and Health." Ph.D. dissertation, University of Chicago, 1996.
- Lee, Chulhee. "Socioeconomic Background, Disease and Mortality among Union Army Recruits: Implications for Economic and Demographic History." *Explorations in Economic History*, January 1997, 34(4), pp. 27-55.
- Moen, Jon Roger. "Essays on the Labor Force and Labor Force Participation Rates: The United States from 1860 through 1950." Ph.D. dissertation, University of Chicago, 1987.
- Waalder, Hans Th. "Height, Weight, and Mortality: The Norwegian Experience." *Acta Medica Scandinavica*, 1984, 679 (Supplement), pp. 1-51.
- Wilson, Sven and Rudberg, Mark. "The Epidemiology of Congestive Heart Failure in the Pre-Antibiotic Era: Evidence from a Cohort of Civil War Veterans." Unpublished manuscript presented at the Conference on the Aging of Union Army Veterans, National Opinion Research Center (NORC), Chicago, IL, 25-27 October 1996.

The Rise of the Welfare State and Labor-Force Participation of Older Males: Evidence from the Pre-Social Security Era

By CHULHEE LEE*

The labor-force participation rate (LFPR) of older males in the United States fell dramatically over the last five decades. Nearly half of men aged 65 and over participated in the labor market in 1945. Today, only 15 percent of males in the same age group are working. The creation and expansion of various public welfare and social-insurance programs for old-age security has often been acknowledged as an important cause of the decline in the labor-market activity of the elderly. Numerous studies have examined how such programs, especially Social Security, affected retirement behaviors of aged males on various empirical grounds.¹ Unfortunately, the results of these studies are sharply divided over what proportion of the decreased LFPR of older males can be attributed to the increase in the coverage and benefits of the government income-transfer programs.

The objective of this paper is to provide insight into how the development of public welfare programs contributed to the decline in the LFPR of the elderly by examining the pattern of change in the LFPR of older males prior to 1940, when few old-age-security aids were offered by the government. I will show that the labor-market activity of older men greatly declined between 1880 and 1940, prior to the creation of major public insurance programs, and that most of the fall in the LFPR of older males during those six decades cannot be explained by the effects of the two major government transfer programs for the elderly introduced prior to 1940, namely, the Union

Army pension and Old Age Assistance (OAA) programs. This result implies that the rise of the welfare state may not be the main cause of the secular decline in labor-force participation of older males.

This study is based on samples of several of the primary data sources that were collected and linked as part of the project titled "Early Indicators of Later Work Level, Disease, and Death." The empirical evidence given below is largely drawn from a sample of 15,149 Union Army recruits who enlisted in companies organized in Ohio, Pennsylvania, and New York. In particular, the patterns of transitions in labor-force status at older ages are examined using a longitudinal sample of 2,259 aged veterans who were linked to both the 1900 and 1910 censuses.²

I. The Labor-Force Participation Rate of Older Men, 1880-1940

Economists have been at odds over the trend in the LFPR of older males before 1940. The conventional belief was that the LFPR among males aged 65 and over remained relatively high until the last decade of the 19th century and began to decline thereafter (John D. Durand, 1948; Clarence Long, 1958; Jon R. Moen, 1987). Roger L. Ransom and Richard Sutch (1986) have challenged this view, suggesting that the LFPR of males aged 60 and over was stable between 1870 and 1937. The main source of the disagreement on the LFPR is how to classify the men who were unemployed for a prolonged period during the census year. Ransom and Sutch (1986) removed individuals reporting six or more months of unemployment (the long-term unemployed, hereafter) from the calculation of the labor

* Department of Economics, Binghamton University, Binghamton, NY 13902-6000. I thank Dora L. Costa for helpful comments. The financial support from The Center for Population Economics, from NIH (PO1 AG 10120), and from NSF (SBR 9114981) is gratefully acknowledged.

¹ See Alan B. Krueger and Jörn-Steffan Pischke (1992) for a recent survey.

² See Lee (1998) for the characteristics and representativeness of the sample.

force, considering them "permanently unemployed." Moen (1987) and Robert A. Margo (1993), on the other hand, contend that the long-term unemployed should not be excluded from the labor force.

To judge the validity of these definitions of the labor force, I examined the patterns of the transitions in labor-force status for the following three groups of aged men in the longitudinal sample of Union Army veterans: (i) those who were employed for the full year or unemployed for less than six months (EMP), (ii) those who were unemployed for six months or longer (UNEMP), and (iii) retirees (RET).³ It turns out that the odds of participation 10 years later for the UNEMP sample is in the middle of those for EMP and RET, holding personal characteristics constant. I also found that UNEMP individuals were much more responsive to changes in pension income and health in making retirement decisions than were EMP individuals. On the other hand, RET individuals were not affected at all by variations in income or health in making labor-force participation decisions.⁴

These patterns of labor-force transitions suggest that the older long-term unemployed were probably at the marginal position in the labor force. Only a small deterioration in productivity or modest increase in nonlabor income would have prompted retirement among them. However, the long-term unemployed are clearly distinct from the retired. They were much more likely to be gainfully employed 10 years later than the retired, and their retirement decisions closely followed the prediction of economic theories of labor-force participation (negative income effect and positive substitution effect), while those of the retired did not.

The above evidence suggests that the long-term unemployed should have comprised a distinct state from nonparticipation, support-

ing the conventional definition of the labor force. However, their higher propensity to retire within the next 10 years and greater responsiveness to changes in income and health compared to those who were employed for the full year indicate that a part of them may have been practically out of the labor force. Judging by the comparison of probabilities of participation 10 years later for the three groups, the long-term unemployed behaved as if they were a 50–50 mix of the employed and the retired.

Based on the above result, I estimate the LFPR of older men for 1880, 1900, and 1910, excluding half of the long-term unemployed from the labor force. One problem is that the incidence of long-term unemployment among aged men (LU, hereafter) for 1880 is not readily available from the micro census data. In estimation, I employ an assumption that LU fell between 1880 and 1900 at the same rate as it did between 1900 and 1910. According to a number of local unemployment statistics, the prevalence of long-term unemployment should not have been much different between 1880 and 1900. Moreover, a sharp decline in LU between 1900 and 1910 was partly due to changes in the definition of unemployment and selection of persons to report the number of weeks of unemployment (Moen, 1994). Under the above assumption, therefore, an upper-bound estimate of LU and a lower bound of the LFPR for 1880 would result. Even for the lower-bound LFPR, a downward tendency is apparent for the period between 1880 and 1910. The LFPR among men aged 65 and over is 70 percent in 1880, 62 percent in 1900, and 58 percent in 1910. The fall in the LFPR among men age 65 and over prior to 1940 accounts for half of the entire decline between 1880 and 1990. This result suggests that the secular decline in the labor-force participation of older males started as early as the end of the 19th century, as the conventional view proposes.

II. Effect of the Union Army Pensions

Dora L. Costa (1995) found that Union Army pensions had a strong positive effect on the probability of retirement of individual pensioners. As Moen (1987) reported, therefore, the decline in the LFPR of older males in the

³ Following Ransom and Sutch (1986) and Moen (1987), I classify as retirees those whose occupation was recorded as "retired," blank, a nonoccupational title (such as "invalid," "live on pension," or "insane"), "capitalist," or "landlord," as well as inmates of institutions.

⁴ See Lee (1998) for more detailed method and results of this regression analysis.

late 19th and early 20th centuries may have been produced in part by the expansion of the Union Army pension program. There is little evidence on the magnitude of the retirement effect of Union Army pensions at the aggregate level, however. As far as the age group of 65 and over is concerned, the effect of Union Army pensions on the LFPR of all males, if any, should have been particularly strong during the first decade of this century. Though the number of pensioners increased rapidly after 1890, the majority of veterans were younger than 65 by the turn of the century. By 1910, however, the vast majority of pensioners were 65 and older. For this reason, the number of pensioners aged 65 and older more than doubled between 1900 and 1910, while the total number of pensioners decreased during the same period. After 1910, on the other hand, the number of pensioners in that age group decreased sharply as many of them died. Accordingly, the analysis below focuses on the period 1900–1910.

I decompose the decline in the LFPR of men 65 and older between 1900 and 1910 into portions attributable to several factors to examine how Union Army pensions affected the trend in the LFPR of older males. The LFPR of males in an age group j may be presented as the weighted average of the LFPR's of Union Army pensioners (P_j^U) and nonpensioners (P_j^N), where corresponding weights are the fractions of pensioners (ϕ_j) and nonpensioners ($1 - \phi_j$) in that age group. Here, I divide males 65 and older into four age groups: 65–69, 70–74, 75–79, and 80 and older. The LFPR of all males 65 and older (P^A) can be given as the weighted average of the LFPR's of the four age groups as below, where ω_j denotes the proportion of men 65 and older in age group j :

$$(1) \quad P^A = \sum \omega_j [\phi_j P_j^U + (1 - \phi_j) P_j^N].$$

By differencing the above equation, one obtains

$$(2) \quad \begin{aligned} \Delta P^A = & \sum \omega_j \phi_j \Delta P_j^U \\ & + \sum \omega_j (1 - \phi_j) \Delta P_j^N \\ & + \sum \omega_j (P_j^U - P_j^N) \Delta \phi_j \\ & + \sum P_j^A \Delta \omega_j + \varepsilon. \end{aligned}$$

The first and second terms on the right-hand side of equation (2) show the change in the LFPR of all males 65 and older produced by changes in the LFPR within the categories of pensioners and nonpensioners. The third term represents the size of the effect of a change in the fraction of pensioners, which depends on the difference in the LFPR between pensioners and nonpensioners. The remaining two terms indicate, respectively, the effect of a change in age structure and the residual.

I estimate the number of pensioners in each age group using the total number of Union Army pensioners reported in Moen (1987) and the age distribution of pensioners in the veteran sample. The number of nonpensioners in an age group is calculated by subtracting the estimated number of pensioners from the total number of the male population in that age group. For the LFPR of pensioners in each age group, I use the participation rate of veterans in the Union Army sample who received pension benefits. The LFPR of nonpensioners in each age group is indirectly estimated based on the LFPR's of all males and pensioners, and the proportion of pensioners in that age group.⁵

The results, reported in Table 1, suggest that the effect of change in the coverage of the Union Army pension program accounts for only 16 percent of the total decline in the LFPR of men 65 and older between 1900 and 1910. The fall in the LFPR among nonpensioners explains nearly half of the LFPR decline among all older males. Another 30 percent is accounted for by the decline in the labor-market activity among pensioners. Even if one assumes that this 30 percent of the decline is entirely attributable to a rise in pension benefits, which is unlikely, less than half of the decline in the LFPR is explained by Union Army pension during the decade in which its effect should have been at the peak.

III. Effect of the Old-Age Assistance Program

According to the conventional estimates (Long, 1958), the LFPR of males aged 65 and

⁵ A more detailed description of the method of estimation and data used in the computation are available from the author upon request.

TABLE 1—A DECOMPOSITION OF THE DECLINE
IN THE LFPR OF MEN 65 AND OLDER, 1900–1910

Variable	Estimate
$\sum \omega_j \phi_j \Delta P_j^U$	0.019 (30.2)
$\sum \omega_j (1 - \phi_j) \Delta P_j^N$	0.028 (44.4)
$\sum \omega_j (P_j^U - P_j^N) \Delta \phi_j$	0.010 (15.9)
$\pi P_j^A \Delta \omega_j$	0.003 (4.8)
ε	0.003 (4.7)
ΔP^A	0.063 (100.0)

Notes: In parenthesis is the percentage of the decline in the LFPR of all men 65 and over between 1900 and 1910 explained by each variable.

over fell from 58 percent in 1930 to 43.5 percent in 1940. This sharp decline in the participation of the aged during this period has often been attributed to the effects of a series of government programs aimed at discouraging the labor supply of older workers, especially the Old Age Assistance (OAA) program, the first major federal–state transfer program to the nonveteran elderly (Donald O. Parsons, 1991; Leora F. Friedberg, 1996). According to Parsons (1991), half of the decline in the LFPR of men 65 and older between 1930 and 1950 can be attributed to the introduction of OAA. This implies that about 20 percent of the total decline in the LFPR of aged males between 1880 and 1940 can be accounted for by OAA.

The pure contribution of OAA to the decline in the LFPR of older males between 1930 and 1940 may have been smaller than suggested by these studies because the strong retirement effect of OAA could have resulted in part from the poor employment opportunities for aged workers during the 1930's. It was suggested above that the long-term unemployed among aged veterans were much more sensitive to a change in income or health in making retirement decisions than were those who were employed for the full year. This result implies that a poor labor-market condition for aged workers could have augmented the retirement effect of income. That is, the OAA benefits could have exerted an even stronger effect on retirement decisions be-

cause of the poor labor-market prospect.⁶ It should be noted that older workers reentered the labor force or suspended their exit temporarily during the period 1940–1950 when the economy had recovered from the Depression, despite the widespread extension of Social Security and private pension plans. It is doubtful, therefore, that OAA could have stimulated retirement under a fair economic condition as much as it did during the Great Depression.

IV. Concluding Remarks

The LFPR of older males started to decline in the late 19th century, long before the development of major public welfare programs for the elderly. Even with a lower-bound estimate of the LFPR for 1880, nearly half of the decline in the LFPR of men 65 and older occurred prior to 1940. Moreover, the two major public income-transfer programs for the elderly during those periods, namely the Union Army pension and OAA programs, appear to be of secondary importance in explaining the trend toward declining labor-market activity of older males. This evidence tends to support the view that the labor-force participation of older males would have declined over the last century even without the rise of public welfare and social insurance programs.

It is widely accepted that increased retirement incomes should be one of the most important causes of the long-term decline in the labor-market activity of older males. Costa (1995), for example, reported that 90 percent of the decline in labor-force participation rates of men older than 64 between 1900 and 1930 could be attributed to secularly rising incomes. A further question arises from this view: what were the main sources of retirement incomes? The result of this study suggests that individuals should have financed retirement relying largely upon private resources, such as savings

⁶ Indeed, Parsons (1991) based his inference on the regression coefficient for the interaction term between the average OAA benefit and the share of workers who earned below the poverty level in each state. The second variable might capture the variation in labor-market condition across states.

in various forms and family supports, rather than income transfers from younger generations forced by the government.

REFERENCES

- Costa, Dora L. "Pensions and Retirement: Evidence from Union Army Veterans." *Quarterly Journal of Economics*, May 1995, 110(2), pp. 367-97.
- Durand, John D. *The labor force in the United States, 1890-1960*. New York: Social Science Research Council, 1948.
- Friedberg, Leora F. "The Effect of Government Programs on the Labor Supply of the Elderly." Ph.D. dissertation, Massachusetts Institute Technology, 1996.
- Krueger, Alan B. and Pischke, Jörn-Steffen. "The Effect of Social Security on Labor Supply: A Cohort Analysis of the Notch Generation." *Journal of Labor Economics*, October 1992, 10(4), pp. 412-37.
- Lee, Chulhee. "Long-Term Unemployment and Retirement in Early-Twentieth-Century America." *Journal of Economic History*, 1998 (forthcoming).
- Long, Clarence. *The labor force under changing income and employment*. Princeton, NJ: Princeton University Press, 1958.
- Margo, Robert A. "The Labor Force Participation of Older Americans in 1900: Further Results." *Explorations in Economic History*, October 1993, 30(4), pp. 409-23.
- Moen, Jon R. "Essays on the Labor Force and Labor Force Participation Rates: The United States from 1860 through 1950." Ph.D. dissertation, University of Chicago, 1987.
- . "The Unemployment and Retirement of Older Men." *Historical Methods*, Winter 1994, 27(1), pp. 40-46.
- Parsons, Donald O. "Male Retirement Behavior in the United States, 1930-1950." *Journal of Economic History*, September 1991, 51(3), pp. 657-74.
- Ransom, Roger L. and Sutch, Richard. "The Labor of Older Americans: Retirement of Men on and off the Job, 1870-1937." *Journal of Economic History*, March 1986, 46(1), pp. 1-30.

Secular Trends in the Determinants of Disability Benefits

By SVEN E. WILSON AND LOUIS L. NGUYEN*

A major justification for devoting resources to the study of public health is the potential to answer questions about the burden of poor health, both in terms of the total burden faced by individuals and the burden placed upon publicly funded social insurance programs. Indeed, the adequate provision of social insurance programs is one of the key policy issues of our day. A potentially fruitful approach in undertaking this effort is to investigate the effects of specific chronic diseases and injuries upon program participation and benefit levels. Ideally, we would like to know something about the total economic costs of individual diseases using theoretically sound willingness-to-pay measures. In practice, however, willingness-to-pay measures cannot be estimated with most available health data.

Though this paper cannot pin down anything as ambitious as the total economic burden of disease, it does address the narrower but still important question of what is the burden of chronic illness upon Social Security Disability Insurance (SSDI) payments, and it documents how that burden has shifted between different disease groups over the past century. Furthermore, it addresses, at least to a limited extent, the profound intellectual question of what determines disability and how biomedical, economic, social, and institutional factors determine whether an individual will be disabled.

In this paper we begin an exploration of newly collected data on the health conditions and disability benefits of Union Army Veterans¹ and make comparisons to recipients of

disability benefits in more recent times. We find two main results. The first is that there has been a significant shift in the types of diseases that lead to disability, both in terms of prevalence rates and benefit levels. The second is more surprising: the disabled in modern times generally have a greater number of chronic illnesses than did disabled Union Army veterans, even those who were severely disabled. This result implies a way of thinking about disease and disability that deserves more research attention. In short, prior to the advent of modern medicine and the concurrent reductions in the physical demands of work, people became disabled not because they had *numerous* chronic illnesses (i.e., high rates of co-morbidity), but because individual conditions (even ones as simple as hernias or hemorrhoids) had much more severely debilitating effects on health and upon the capacity to work than those same conditions do today.

I. Data

Data for this study come from two major sources. The first data source includes extensive individual-level data relating to military service, health, occupation, residence, and a host of other socioeconomic variables across the lifespan of sample Union Army veterans of the U.S. Civil War. When the collection is complete it will contain a longitudinal, random sample consisting of several thousand variables on approximately 36,000 individuals that were mustered into the Union Army during the Civil War (referred to here as the UA sample). This sample has been shown to be representative of the white, male, Northern population in the late 19th century.²

The second major data source is the New Beneficiary Survey (NBS). In 1982, the Social Security Administration surveyed

* Wilson: Departments of Political Science and Economics, Brigham Young University, 762 SWKT, Provo, Utah 84602; Nguyen: Department of Surgery, Children's Hospital, Harvard University, Enders 1050, 300 Longwood Avenue, Boston, MA 02115.

¹ Data come from "Early Indicators of Later Work Levels, Disease, and Death" (Robert W. Fogel and Larry T. Wimmer, 1992). All data used here are being prepared for deposit at the Inter-university Consortium for Political and Social Research (ICPSR), Ann Arbor, MI.

² For evidence, see Fogel (1993).

individuals who had recently begun receiving benefits. The NBS is valuable because it contains extensive information on both income sources and amounts (including SSDI) and data on specific chronic diseases and impairments.

Making comparisons between these two data sources is extremely challenging for a number of reasons. First the state of medical knowledge is much different between the two time periods. Second, medical data in the UA sample come from physical exams, while the NBS health data is self-reported. Third, the eligibility requirements for the Union Army pension and SSDI were much different, as was the determination of benefits. Disability benefits in the UA data were determined by severity, while under SSDI they are determined primarily by Social Security earnings history. Finally, because we are talking about people living a century apart, many other differences exist between the populations that can never be completely controlled for. This is the same challenge that has always existed in using historical data to make comparisons to modern populations.

In order to make the data as comparable as possible, several restrictions were made in including observations for analysis. Both samples are restricted to "new beneficiaries" of disability benefits. In the NBS, all individuals are new recipients of SSDI, while in the UA sample, we restrict analysis to those who were first pensioned in 1891–1892. This period was chosen because it follows the liberalization of the pension laws in 1890 which allowed, for the first time, disability benefits to be awarded for conditions not related to wartime experience. Both samples were also restricted to males aged 45–59. The final NBS subsample includes 1,569 individuals, while the UA subsample contains 1,410 individuals mustered into the Union Army in 15 different states and the District of Columbia.

The bulk of the health information in the UA sample comes from records of the Pension Bureau, referred to as "Surgeons' Certificates." After applying for disability benefits, the claimant was required to submit to a detailed physical examination by a government-appointed board of three physicians. These exams were very thorough and included an in-

vestigation of the particular disabilities claimed by the veteran. The examining physicians noted physical-exam findings and recommendations for pension benefits on the certificate. Previous studies using the UA data on health were from a small pilot sample in which the Surgeons' Certificates were collected in a very crude fashion. The data used here, in contrast, contain roughly two-thirds of the final sample size, and the collection process of the medical data has been dramatically improved.

Data from the Surgeons' Certificates are organized into major disease categories that reflect the categorization typically given by the examining physicians. Each of these categories is given a "rating," which corresponds to the dollar amount for the categories. For example, if the claimant was found to have heart disease and "rheumatism," the dollar amount awarded was divided between these two conditions according to their respective severity. Some of the disease categories correspond obviously to major body systems (cardiovascular, respiratory, genitourinary, etc.). Other categories are narrower, but they were important in terms of the disability rating system. For instance, there are categories for chronic diarrhea, varicose veins, and hernias because the examining physicians would frequently give dollar ratings attributed to these individual conditions. The disability rating is used as the fundamental criterion of whether disease exists within a given category. There are, of course, extensive additional data collected from the certificates that can be used to determine the presence of disease. It should be noted that this is a difficult and subjective process using currently nonstandardized values. Further explorations into making differentiated diagnoses may alter somewhat the results given here.³

³ Because of space restrictions, a detailed discussion of how determinations of disease are made is not possible. An appendix outlining the methodology used in making these determinations, as well as more detail on disease categories in both the Union Army sample and the New Beneficiary Survey, is available from the authors upon request.

TABLE 1—NEW BENEFICIARIES OF SSDI (1982) AND NEW BENEFICIARIES OF THE UNION ARMY PENSION (1891–1892)

Disease category	SSDI recipients (percentage)	Union Army veterans by disability status (percentage)		
		Mild	Moderate	Severe
Cardiovascular	70.5	25.2	40.3	48.3
Arthritis	60.7	64.9	65.9	68.0
Injury	50.9	24.4	28.8	32.0
Mental/emotional	38.1	2.0	3.3	4.4
Gastrointestinal	37.4	44.1	56.1	61.6
Eye	37.2	13.0	15.3	20.7
Respiratory	30.4	19.1	18.4	31.3
Ear	20.9	4.5	7.2	25.2
Central nervous system	19.6	8.0	10.3	16.3
Cancer	6.8	0.5	1.1	2.0
Other	NA	11.3	18.4	22.1
N:	1,569	644	472	294

Notes: Both samples include males aged 45–59. The SSDI sample has been reweighted to match the age distribution of the Union Army Veterans' sample. Information on disease classification is provided in a methodological appendix available from the authors upon request. See text for discussion of disability status.

II. Comparing the Health of the Disabled Across Time

While we do not make a direct comparison between the overall rates of disability-program participation between the two time periods, it is clear that the program participation rate was much higher for the Union Army pension than for SSDI. While the exact rate of participation among veterans is not known, the UA sample indicates that by 1890, between 26 percent and 47 percent (depending on mortality assumptions) of veterans had already enrolled in the program. In contrast, roughly 3 percent of the working-age population receives SSDI.

What we are able to calculate here is the proportion of individuals in each sample who have specific chronic conditions. Table 1 makes this comparison using a slight reclassification of the disease categories in each sample.⁴ The first column shows the percent-

age of sample respondents in the NBS who report specific chronic conditions.⁵ The remaining columns give the prevalence rates for the Union Army veterans according to the overall disability status of the veterans. The designations of mild, moderate, and severe are somewhat arbitrary, but natural breaking points occur given the nature of the Union Army pension system during this time period. "Severe" disability occurs if the examining physicians recommend a pension of at least \$18 per month (which is a "total 3rd grade" pension and is considered equivalent to the loss of a hand or foot). "Moderate" disability is given to those who are recommended to receive \$12–17 per month (\$12 is the maximum allowable under the 1890 law). "Mild" disability is any recommended amount less than \$12 per month.

Given that the Union Army program had much higher rates of participation, it is not surprising that, overall, the disease prevalence rates are higher for the NBS sample. However, even when we restrict the UA sample to those who are severely disabled (21 percent of the sample), we find that the prevalence rates in the NBS are either roughly the same (within 5 percentage points) or higher than the UA rates. The only exception to this is the dramatically higher rate of gastrointestinal disease (hernias, chronic diarrhea, hemorrhoids, diseases of the liver, to name a few) among the UA individuals. The UA sample has somewhat higher rates for arthritis, for respiratory disease, and for diseases of the ear, but the SSDI sample is substantially higher for the disease categories of cardiovascular, injury, mental/emotional, and eye disease. Diseases of the central nervous system and cancer are also modestly higher in the NBS sample.

Issues of sample comparability should lead us to approach these results cautiously, because there are many potential sample biases. One reason the NBS rates may be higher is that these individuals are recent recipients of SSDI, and they may feel an

⁴ This reclassification is also discussed in the methodological appendix (see footnote 3).

⁵ The reported numbers for the SSDI sample have been reweighted to match the age distribution of the Union Army Veteran Sample, though this reweighting has very small effects on the reported results.

incentive to overreport their health conditions given that they have recently been classified as disabled. However, the general scholarly consensus on self-reported health data is that most conditions are underreported, if for no other reason than that they have not been revealed by a physical exam, thereby mitigating the effect of the overreporting. An additional potential bias is that the UA sample may contain many individuals who would have previously been classified as disabled but were not eligible until the change in the law in 1890, whereas the NBS sample includes only those recently disabled. However, since there was little that could be done for these people in terms of medical treatment, the UA sample rates are probably higher than they would be if the sample were to consist solely of the newly disabled, as the NBS does. This bias tends to reinforce, not weaken, our results.

Probably the biggest reason to regard this result as tentative involves differences in diagnostic techniques between the two periods. One advantage of a disease-specific approach is that we can point to particular conditions where this effect should be most pronounced. Cancer is an obvious example, given that most cancers were undetectable and untreatable in the 19th century. Another case is mental/emotional illness. It is also plausible that applying advanced diagnostic techniques in the case of cardiovascular disease would lead us to increase the UA estimate in that category, particularly since hypertension was not understood in the 1890's but is a common diagnosis today.

It is important to note at this point that higher rates of chronic illness among recipients of disability benefits do not imply that the UA individuals were not as sick as modern populations. There are compelling reasons to believe that the physical disabilities of individuals in the 1890's were, on average, much greater than today, even when holding constant the level of chronic disease and injury. For instance, lack of medical treatment caused many conditions that are effectively treated today (such as hernias or varicose veins) to be highly debilitating. This is especially true in light of the fact that the physical demands of the workplace have fallen significantly.

TABLE 2—RELATIVE CONTRIBUTION OF INDIVIDUAL DISEASE CATEGORIES TO DISABILITY BENEFIT LEVELS (PERCENTAGES)

Disease category	Union Army veterans (1891–1892)	SSDI recipients (1982)
Gastrointestinal	25.4	9.3
Arthritis	23.3	15.9
Cardiovascular	12.4	23.3
Injury	10.0	13.9
Respiratory	8.3	8.2
Other	6.0	0.0
Eye	5.8	8.6
Ear	5.1	4.5
Central nervous system	2.8	4.9
Mental/psychological	0.6	9.3
Cancer	0.4	2.2
N:	1,410	1,569

Notes: Both samples include males aged 45–59. The SSDI sample is reweighted to match the age distribution of the Union Army sample. Disease reclassifications are discussed in a methodological appendix available from the authors upon request.

A final comparison between the two periods concerns not the prevalence of conditions among the disabled, but the relative contributions that chronic conditions make to the monetary value of disability benefits. In the Union Army sample, dollar amounts were specifically attributed to different conditions, but in the NBS, only crude estimates of relative burden can be made. This task is undertaken here by dividing the total monthly benefit equally across the conditions reported by the respondent. Thus if the total benefit is \$600 and the respondent reports having a respiratory condition and arthritis, then the \$300 is attributed to each condition.

The relative budgetary burdens of the different conditions are classified in Table 2 (the observations from the NBS sample have been reweighted to reflect the age distribution in the Union Army sample). Though similarities exist, there is a noticeable shift in the distribution of disease burden. Gastrointestinal disease (which includes mostly conditions that are easily correctable today) and arthritis together account for close to half the burden of disease of new beneficiaries in the 1891–1892 period, but just over 25 percent in the modern period. Much of the difference has been made up in

terms of cardiovascular disease and mental/emotional conditions. These shifts, though tentative given the nature of the data, are consistent with the medical, economic, social, and institutional factors that have occurred over the last century.

III. Future Directions

The economic burden of disease is represented, in part, by the payment of disability benefits. Thus, the relative burden of different chronic health conditions is an important variable that affects policy choices such as the allocation of medical research funds or the provision of disability insurance. The lesson of this research is that accurate forecasts of disease prevalence rates are not sufficient for making disability policy.⁶ We must understand, as well, the variety of continually

changing factors that cause a particular condition to be debilitating and whether or not that debilitation leads to participation in available disability programs.

REFERENCES

- Fogel, Robert W. "New Sources and New Techniques for the Study of Secular Trends in Nutritional Status, Health, Mortality and the Process of Aging." *Historical Methods*, Winter 1993, 26(1), pp. 5-43.
- Fogel, Robert W. and Wimmer, Larry T. "Early Indicators of Later Work Levels, Disease, and Death." Working Paper Series on Historical Factors in Long Run Growth, National Bureau of Economic Research (Cambridge, MA) Historical Working Paper No. 38, June 1992.
- Manton, Kenneth G.; Corder, Larry and Stallard, Eric. "Chronic Disability Trends in Elderly United States Populations: 1982-1994." *Proceedings of the National Academy of Science USA*, 18 March 1997, 94(6), pp. 2593-98.
- Waidmann, Timothy; Bound, John and Schoenbaum, Michael. "The Illusion of Failure: Trends in the Self-Reported Health of the U.S. Elderly." National Bureau of Economic Research (Cambridge, MA) Working Paper No. 5017, February 1995.

⁶ The link between specific chronic disease and disability is missing, for instance, in the recent finding by Kenneth G. Manton et al. (1997) that chronic disability rates among the U.S. elderly (age 65+) fell between the years of 1982 and 1994. Though there is mounting evidence, such as that in Timothy Waidmann et al. (1995), that rates of chronic illness are falling as well, little is known about how changes in the relationship between disease and disability may have affected the disability trends.

The Evolution of Retirement: Summary of a Research Project

By DORA L. COSTA*

We are not the first generation faced with rising retirement rates and an increasing proportion of elderly. In 1880, less than 3 percent of the U.S. population was older than 64, and this figure has increased steadily over the course of this century. Retirement rates of older men, both in the United States and in Europe, have been rising for more than a century as well. In 1880, the majority of men older than 64 toiled in the labor force. The fraction in the labor force has fallen steadily, and continuously, so that those men left in the labor force today are in the minority (see Fig. 1).

The trend observed in Figure 1 suggests that the past can inform our predictions of the future course of retirement rates. Seventy percent of the decline in the labor-force participation rates of men age 65 or older occurred before 1960. Furthermore, the trends across countries are remarkably similar, suggesting that characteristics of state or private pensions that are specific to an individual country cannot explain the long-term rise in retirement rates.

Researchers who have sought to examine the origins of retirement have been stymied by the lack of data. The most common explanation for increasing retirement rates has been rising incomes, whether in the form of state or private pensions or personal savings. An analysis of retirement therefore requires information on retirement status; demographic characteristics such as age; health; a proxy for the opportunity cost of not working such as

forgone income or occupation; and retirement income such as pension amount. These are very strict data requirements. One of the few sources of information on the elderly of the past, the census, allows one to relate retirement status only to demographic characteristics, not to wealth. Fortunately, with enough time and money, it is possible to create a longitudinal data set that meets these requirements by linking census records to the records generated by the first major pension program in the United States, that serving Union Army veterans of the Civil War.

This paper summarizes the results of a research project (Costa, 1998) on the evolution of retirement that uses Civil War records to compare the retirement experience of the cohort that reached age 65 at the beginning of the 20th century with that of later cohorts.

I. Income and Retirement

Although evidence on the impact of assets, private pensions, and Social Security on retirement rates is mixed, many studies find that these factors have a relatively small effect on retirement rates. For example, Patricia M. Anderson et al. (1997) conclude that changes in pensions and social security from the late 1960's into the 1980's account for at most a quarter of the trend toward earlier retirement for those in their early sixties and cannot account for any of the increase in retirement rates among those older than 64. But, this need not imply that rising retirement incomes have had little impact on long-run retirement rates. The applicability of cross-sectional estimates based on data from the 1960's and later to periods outside the sample range is questionable. Retirement rates were already high by 1960, and thus only large benefit increases could have enticed those remaining in the labor force to have withdrawn.

The Union Army records provide a unique opportunity to estimate the impact of a pure income transfer on retirement rates in the

* Department of Economics, Massachusetts Institute of Technology, E52-274C, 50 Memorial Drive, Cambridge, MA 02142-1347, and NBER. This research was funded by NIH grants AG12658 and AG10120. A preliminary release of the data used in this paper is available from Inter-university Consortium for Social and Political Research (Ann Arbor, MI) as study numbers 6836 and 6837. More current releases are available from Julene Bassett, Center for Population Economics, Graduate School of Business, University of Chicago, 1101 E. 58th Street, Chicago, IL 60637.

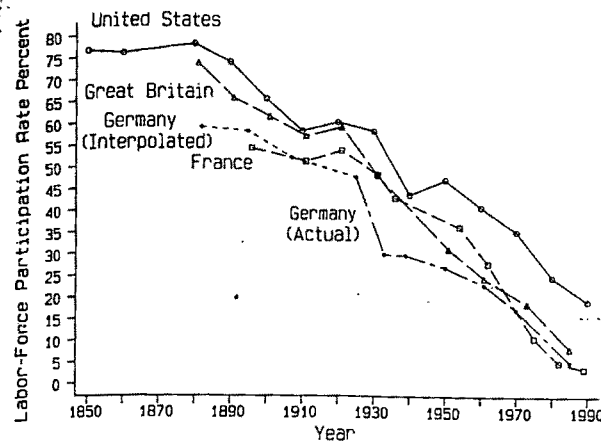


FIGURE 1. LABOR-FORCE PARTICIPATION RATES OF MEN AGE 65 AND OVER, 1850-1990, UNITED STATES, BRITAIN, FRANCE, AND GERMANY

Note: German participation rates for 1882, 1895, and 1907 are based on the participation rates for men age 60-69 and 70 or older.

Source: Costa (1998).

past. The receipt or level of Union Army pensions, which replaced about 30 percent of the income of an unskilled laborer, did not depend on current income or past wages. Rather, their generosity was determined by the pensioner's health. Because the amount received depended not just on the seriousness of his infirmity, but also on whether the veteran could trace his disability to the war and therefore on the often incorrect medical theories of the time, the impact of pensions on labor supply can be disentangled from that of health.

Estimates of elasticities of retirement with respect to pensions suggest that Union Army veterans were very responsive to pension income. In 1900 the elasticity was 0.73, and in 1910 it was 0.46. Recent elasticities are much smaller (see Table 1). The 1910 estimate of the elasticity of retirement implies that 90 percent of the decline in labor-force participation rates of men older than 64 between 1900 and 1930 could be attributed to secularly rising incomes. Leora F. Friedberg's (1996) estimates suggest that secularly rising incomes account for at least half of the decline between 1930 and 1950. Rising incomes may explain none of the 1950-1980 decline.

TABLE 1—ELASTICITIES OF RETIREMENT WITH RESPECT TO RETIREMENT INCOME IN SELECTED STUDIES

Study	Age group	Year	Elasticity
Union Army pensions			
Costa (1998)	median = 56	1990	0.73
Social Security disability			
Donald O. Parsons (1980)	48-62	1969	0.63
Jonathan S. Leonard (1979)	45-54	1972	0.35
Haveman and Wolfe (1984a, b)	45-62	1978	0.21-0.06
John Bound (1989)	45-64	1972/1978	0.16
Union Army pensions			
Costa (1998)	median = 66	1910	0.47
Old Age Assistance			
Friedberg (1996)	66-80	1940/1950	0.25-0.42
Old Age and Survivors Insurance			
Michael D. Hurd and Michael J. Boskin (1984)	60-64	early 1970's	0.71
Jerry A. Hausman and David Wise (1985)	58-63 in 1969	1969-1979	0.23
Alan B. Krueger and John-Steffen Pischke (1992)	60-68	1976-1988	~0
Assets			
Hausman and Wise (1985)	58-63 in 1969	1969-1979	~0

Note: The elasticities given for Friedberg (1996), Hurd and Boskin (1984), and Hausman and Wise (1985) were estimated using the information provided by the authors.

Several factors could account for the falling income elasticity of retirement. Workers may now be less responsive to changes in transfer income because they are no longer close to subsistence levels; instead, they reach retirement age with enough to satisfy their consumption needs. Alternatively, by establishing age 65 (and later age 62) as an "official" retirement age, Social Security may have led individuals to want to retire at that age and therefore reduced the effect of income on the work decision. Finally, retirement has become more attractive because men are less circumscribed in their choice of leisure-time activities. Mass tourism and mass entertainment have increased the variety of recreational activities and lowered their price.

II. Other Explanations

Other explanations for the rise of retirement have focused on increases in the duration of unemployment, sectoral shifts in the economy, and worsening average health. But a comparison of recent with early cohorts, including Union Army veterans, suggests that the burden of chronic disease has fallen. Between 1910 and 1983 the prevalence of heart disease among men older than 64 fell from 75 percent to 40 percent, that of musculoskeletal disease from 68 percent to 48 percent, and that of respiratory disease from 42 percent to 30 percent. Between 1935 and 1992 rates of blindness fell by about one-third. The elderly have benefited from advances in medical technology, lessened occupational hazards, and better early-life conditions. At the same time, health has become less important to the retirement decision. Because chronic conditions can now be better controlled and because physical job requirements have been reduced, those in poor health are more likely to participate in the labor force now than in 1900 or 1910 (relative to those in good health). Age 65 may therefore no longer be as appropriate a demarcation of old age as it was in the first half of the century, when the health of the typical 65-year-old was very poor.

Declines in part-time work, nonfarm self-employment, and farming cannot explain the rise of retirement since 1880 either. An examination of census data reveals that the proportion of 65–74-year-old employees working part-time rose from 15 percent in 1940 to 47 percent in 1990. The fraction of the labor force that is self-employed has fallen, but only since the 1960's have older self-employed workers been more likely to remain in the labor force than wage and salary workers. The lower retirement rates of farmers relative to nonfarmers are also a recent phenomenon. The longitudinal information on occupation and residence in the Union Army records shows that in 1900 and in 1910 farmers were no less likely to retire than nonfarmers and that upon retirement farmers usually moved to a nearby town, thus exiting the farm population.

The increased duration of unemployment spells accounts for up to one-fifth of the increase in retirement rates of men older than 64

since 1900. Census data suggest that unemployment within the state of residence has a substantial effect upon the retirement of men older than 64 in 1900 and upon men aged 50–64 in 1980. But the unemployed would not have been able to retire unless they had income sources other than wages. In 1910, men faced with high unemployment in their state of residence were more likely to leave the labor force if they were Union Army veterans (and hence eligible for a pension) than if they were nonveterans.

III. The Retirement Lifestyle

For many men at the beginning of the century, retirement brought with it dependence upon children. Close to half of retired men older than 64 in 1880 were living in the households of children or other relatives. Estimates of the impact of Union Army pensions on the probability of living with family members suggest that at the beginning of the century retired men older than 64 would have preferred to remain independent of their families; the majority simply could not afford to do so.

By mid-century fewer than 20 percent of retired men were dependent upon their children for support, and some of these men had migrated to Florida or California to enjoy their retirement in warmer weather. Even if they had not migrated, they were able to enjoy many more recreational amenities than their predecessors. In addition to socializing with other retirees or reading, they could spend their time listening to the radio, going to movie theaters, or touring in their own cars. In fact, by the mid-1930's the purchase of recreational goods was no longer concentrated largely among the well-off, but was fairly egalitarian across all income groups.

The fraction of 65-year-old retirees citing a preference for leisure as their main motivation for leaving the labor force rose from 3 percent in 1941 (Edna C. Wentworth, 1945) to 48 percent in 1982 (Sally R. Sherman, 1985). Upon retirement, many couples are now increasing both the frequency and length of travel (Dean W. Morse and Susan H. Gray, 1980 p. 60). Only 5 percent of retired men older than 64 are now living in the households of their children. Evidence from recent data (see Douglas

Wolf [1994] for a review of the literature) implies that changes in income have relatively little effect on living arrangements. The few retired men living with their children may have special needs or tastes. In addition, not only are incomes now much higher than in the past, but the growth of retirement communities in low-cost living areas, the declining price of transport and of communication with family members, and the rise in private and state social support services, among other factors, have lowered the price of living alone. This in turn may have increased the attractiveness of retirement.

Retirement has become a meaningful concept for women as well. In the past, relatively few women devoted their prime years to market work. Women who entered the labor force early in their lives withdrew to work in the home, and if they reentered once their children reached adulthood they left the labor force again in their late forties or early fifties. The important life event that they faced in old age was widowhood, accompanied by dependence upon their children. Although widowhood still characterizes women's experience of old age, Social Security has reduced widows' dependence upon their children (Costa, 1997).

IV. Looking to the Future

The elderly have financed part of their retirement through public monies. First it was through Union Army pensions; then in the late 1920's and early 1930's, many states provided pensions to the needy aged. These pensions were later replaced by Social Security Old Age Assistance and Old Age Insurance. The availability of revenue resources and increasingly well-organized elderly pressure groups spurred the growth of old-age programs. As the population ages, the elderly may become an even more powerful political force pushing for program expansion. But, as their numbers rise, it will become increasingly harder for the young to finance a lengthy retirement for the old.

Changes to the Social Security system designed to ease the financial burden on the young, such as increasing the age at which full benefits can be received, have already been implemented, and more are likely. Will these

changes be enough to slow down or reverse the trend toward earlier retirement? Figure 1 showed that, although historically there have been increases in labor-force participation rates (e.g., between 1910 and 1920 and between 1940 and 1950), the overwhelming trend has been toward falling participation rates. Table 1 demonstrated that income has exerted less and less of an effect on the retirement decision of succeeding cohorts. If this insensitivity to income arises from social norms, perhaps established by Social Security, then a reversal certainly is possible. But part of this insensitivity may arise from the relatively high wealth levels of today's retirees, from their ability to maintain their standard of living by migrating to low-cost areas, and from the affordability of mass tourism and mass entertainment, both of which enable the elderly to pursue "the good life." The expenditure elasticity of recreational goods has declined sharply since the 1880's (a decline that coincides with that in the income elasticity of retirement), suggesting that income has become less important to the enjoyment of leisure. If incomes continue to rise as economic growth progresses and if leisure time activities continue to be relatively inexpensive and enticing, then the rise of retirement is unlikely to reverse.

Without a reversal in the rise of retirement, payroll taxes will have to increase to resolve the fiscal problems facing Social Security. These fiscal problems are likely to be compounded by the increasing longevity of the elderly population. The Union Army records show that the American population has been growing healthier since the end of the 19th century. Trends in adult height and early life conditions suggest that the baby boomers will enjoy a particularly healthy and long-lived old age. Although morbidity improvements may reduce health-care costs, the elderly population will be larger both in terms of absolute numbers and as a proportion of the population. Previous generations responded to an increasing proportion of elderly by instituting old-age pensions. But as the number of years spent in retirement begins to approach one-third of our adult lives and as retirement continues to evolve from a period of dependency to a period of self-realization, the financing of

retirement is likely to rest increasingly with individuals.

REFERENCES

- Anderson, Patricia M.; Gustman, Alan L. and Steinmeier, Thomas L. "Trends in Male Labor Force Participation and Retirement: Some Evidence on the Role of Pensions and Social Security in the 1970's and 1980's." National Bureau of Economic Research (Cambridge, MA) Working Paper No. 6208, October 1997.
- Bound, John. "The Health and Earnings of Rejected Disability Insurance Applicants." *American Economic Review*, June 1989, 79(3), pp. 482-503.
- Costa, Dora L. "A House of Her Own: Old Age Assistance and the Living Arrangements of Older Nonmarried Women." National Bureau of Economic Research (Cambridge, MA) Working Paper No. 6217, October 1997.
- . *The evolution of retirement: An American economic history, 1880-1990*. Chicago: University of Chicago Press, 1998.
- Friedberg, Leora F. "The Effect of Government Programs on the Labor Supply of the Elderly." Ph.D. dissertation, Massachusetts Institute of Technology, 1996.
- Hausman, Jerry A. and Wise, David A. "Social Security, Health Status, and Retirement," in David A. Wise, ed., *Pensions, labor, and individual choice*. Chicago: University of Chicago Press, 1985, pp. 159-91.
- Haveman, Robert H. and Wolfe, Barbara L. "The Decline in Male Labor Force Participation." *Journal of Political Economy*, June 1984a, 92(3), pp. 532-41.
- . "Disability Transfers and Early Retirement: A Causal Relationship?" *Journal of Public Economics*, June 1984b, 24(1), pp. 47-66.
- Hurd, Michael D. and Boskin, Michael J. "Effect of Social Security on Retirement in the Early 1970s." *Quarterly Journal of Economics*, November 1984, 99(4), pp. 767-90.
- Krueger, Alan B. and Pischke, Jörn-Steffen. "The Effect of Social Security on Labor Supply: A Cohort Analysis of the Notch Generation." *Journal of Labor Economics*, October 1992, 10(4), pp. 412-37.
- Leonard, Jonathan S. "The Social Security Disability Program and Labor Force Participation." National Bureau of Economic Research (Cambridge, MA), Working Paper No. 392, August 1979.
- Morse, Dean W. and Gray, Susan H. *Early retirement—Boon or bane? A study of three large corporations*. Montclair, NJ: Allanheld, Osmun, 1980.
- Parsons, Donald O. "The Decline in Male Labor Force Participation." *Journal of Political Economy*, February 1980, 88(1), pp. 117-34.
- Sherman, Sally R. "Reported Reasons Retired Workers Left Their Last Job: Findings From the New Beneficiary Survey." *Social Security Bulletin*, March 1985, 48(3), pp. 22-25.
- Wentworth, Edna C. "Why Beneficiaries Retire." *Social Security Bulletin*, January 1945, 8(1), pp. 16-20.
- Wolf, Douglas A. "The Elderly and Their Kin: Patterns of Availability and Access," in Linda G. Martin and Samuel H. Preston, eds., *Demography of aging*. Washington, DC: National Academy Press, 1994, pp. 146-94.

Uncertain Demographic Futures and Social Security Finances

By RONALD LEE AND SHRIPAD TULJAPURKAR*

Long-run decline of fertility and mortality will lead to secular aging of the U.S. population, punctuated by the retirement of the baby-boom generations in the early 21st century. These changes will severely stress our Social Security system. However, despite their inexorability, there is uncertainty about their timing and extent, and about economic factors which will also influence the system's finances. Projection scenarios based on high, medium, and low assumptions are widely used to assess and describe this uncertainty, but this approach suffers many problems and has no probabilistic interpretation.

In this paper, we build on our earlier stochastic forecasts of the population (Lee and Tuljapurkar, 1994), combined with time-series models of productivity growth and interest rates, to construct stochastic forecasts of the long-run finances of Social Security (OASDI). Our forecasting models do not incorporate economic feedbacks; interest and labor productivity are treated as exogenous. The Congressional Budget Office (1996) has also prepared long-run stochastic federal budget forecasts based on our (1994) stochastic populations, and Martin Holmer (1995) has developed stochastic Social Security projections using a different approach.

I. Forecasts of Fertility, Mortality, and Population

Our stochastic population forecasts are based on prior stochastic forecasts of mortality

(Lee and Lawrence Carter, 1992) and fertility (Lee, 1993), taking net immigration as given at the levels assumed by the Social Security Administration forecasts (Board of Trustees, 1995, 1996), henceforth called SSA. Age-specific death rates, $m(x, t)$ (where x is age and t is time), are modeled as

$$\ln(m(x, t)) = a(x) + b(x)k(t) + \varepsilon(x, t).$$

Here $k(t)$ is an unobserved but estimable index of the intensity of mortality. The model is fit to a matrix of death rates $m(x, t)$, and the resulting estimated time series of $k(t)$ is modeled as a stochastic time series (for the United States, a random walk with drift fits the series well). The probability distribution of $k(t)$ can then be forecast and used to generate probability distributions for future age-specific death rates. In these forecasts, mean life expectancy rises from 76 now to 86 years by 2070, twice the gain in SSA projections, with a 95-percent interval of 81–90 years. Fertility is handled similarly, except that the long-term mean level of fertility is constrained to a specified level, here taken to be 1.9 children per woman, as assumed by SSA. The predicted 95-percent interval converges to 0.7–3.3 children per woman.

Probabilistic forecasts of the population are then constructed through repeated stochastic simulations, with each generating one particular realization or sample path for the population size and age distribution from 1995 to 2070. Probability distributions are computed from the frequency distributions of these sample paths or functions of them. The mean forecast for the old-age dependency ratio in 2070 is 0.47 with a 95-percent interval of 0.26–0.68. The corresponding SSA projection is 0.41 (0.31–0.57).

[†] Discussant: James M. Poterba, Massachusetts Institute of Technology.

* Departments of Demography and Economics, University of California, 2232 Piedmont Ave., Berkeley, CA 94720, and Mountain View Research, 2251 Grant Road, Mountain View, CA 94024, respectively. Lee's research is funded by a grant from NIA, AG11761. Tuljapurkar's research is funded by a grant from NICHD, HD32124.

II. Forecasts of Productivity Growth Rates and Interest Rates

SSA considers alternative assumptions about fertility, mortality, productivity, interest, immigration, inflation, disability uptake, and disability termination. We treat only the first four as stochastic. The growth rate of real productivity and the real interest rate are modeled as undifferenced stochastic time series with long-run means constrained to equal the middle assumptions of SSA (1 percent per year and 2.3 percent, respectively; the fitted models turn out to be independent AR(1) processes). Not differencing and constraining the mean alleviates the serious problems in the use of statistical time-series models for long-term forecasts (wandering means and exploding variances). Productivity is measured as output per labor hour, purged of the effects of changing age-sex composition of the labor force. Interest rates are those on special Treasury bonds issued for Social Security. The farther the trust-fund balance departs from zero, the less defensible is the assumption of exogeneity for productivity growth and interest rates. Our 95-percent interval for productivity growth converges to a width of 0.08, whereas the width of the SSA bracket is 0.01, or only one-eighth as large. The comparison of intervals for interest rates is similar. Given the different meaning of these intervals in the different contexts, however, they are not necessarily inconsistent (see Lee and Tuljapurkar [1998] for details).

III. The Tax and Benefit Schedules

To calculate the tax revenues and benefit payments of the system, we begin with empirical age profiles of 1994 payroll taxes paid and benefits received (calculated from the March Current Population Survey and the Statistical Supplement to the Social Security Bulletin). We then forecast these age schedules for future years conditional on assumptions about policy. In the "current policy" forecasts, we modify the shape and level of future schedules to reflect legislated changes in the normal retirement age in the coming decades. We raise the age schedule of payroll taxes in each future year in proportion to projected productivity.

We alter the age schedule of retirement benefits as a function of productivity growth in a complicated manner reflecting intercohort variations in the average wage when cohorts turn 60, following the benefit rules. Slower productivity growth tilts the age schedule up at older ages; faster growth does the opposite. Our benefit schedule implicitly reflects such factors as the selectivity of mortality by benefit level at higher ages and the increasing proportions of widow(er)s at higher ages, both of which make benefits rise over time for a cohort. However, these features should be functions of the changing level of mortality, but in our simulations they are not. The same is true for survivors' benefits.

IV. Mechanics of the Forecasts

Our forecasts of Social Security finances combine these components in a straightforward way. Along any particular sample path, we have realizations of fertility, mortality, and the corresponding population age distribution for each year, as well as productivity growth and interest rates. We start with an initial trust-fund level as of 1994, an initial population age distribution, and some assumption about the policy regime. Using the procedures just described, we calculate the tax revenues for the first year (multiplying the projected population age distribution times the projected payroll tax schedule, and summing), and add this to the interest received on the fund and to income taxes on benefits. From this total we subtract the benefit payments plus a small amount for administrative costs plus legislated transfers to the railroad retirement fund. The trust-fund level is then augmented by the resulting number, and the whole process is repeated for the next year, up to 2070. We repeat this process for 750 different sample paths and then report the means and frequency distributions of quantities of interest.

V. Forecasts Conditional on Current Policy

We first assume that current policy continues, including legislated changes in the normal retirement age. In this case, the mean fund balance crosses the line of zero reserves in 2026, three years earlier than in SSA forecasts. The

main message, however, is uncertainty about this crossing point: the 95-percent interval includes fund exhaustion as early as 2014, as well as exhaustion as late as 2037, with some sample paths never reaching exhaustion. The level of the reserve fund peaks at \$1.3 trillion (in 1996 dollars, here and throughout) on average, but the 95-percent interval includes \$4.0 trillion and \$0.57 trillion.

If we continue past the date of exhaustion, the forecasted fund debt becomes enormous, with its mean growing to a debt of \$26 trillion by 2070. The 95-percent range spans debts in 2070 of \$6–60 trillion. Alternatively, the median debt in 2070 is three times the payroll, and the upper bound is 12 times the payroll. Obviously, the debt could not actually grow so large without serious consequences for the economy, which would in turn cause changes in interest rates and productivity growth or lead to policy changes in taxes or benefits. Because we have not modeled any economic or policy feedbacks, these forecasts should not be taken at face value. They are based on a premise which they reveal to be false. We do not believe that the trust fund will actually go to zero, let alone to debts of many trillions of dollars. As the fund falls, action will be taken to raise the tax schedule, reduce benefits, delay retirement, further tax benefits, invest the reserves in equities, privatize the system, or in some other way prevent insolvency.

We believe it is more interesting and useful to ask different questions and forecast different quantities. The long-term actuarial balance (LTAB) indicates the immediate tax increase required to equalize the present value of non-interest revenues through 2070 and the present value of costs, with reserves in 2070 equal to one year's benefits. The current payroll tax for OASDI is 12.4 percent. In its "middle" projection, SSA calculates that the LTAB is -2.2 percentage points (Board of Trustees, 1996) so that payroll taxes would have to be raised by 2.2 percentage points to achieve long-term balance. The LTAB is somewhat misleading, due to its 75-year horizon. With a 2.2-percentage-point increase, benefit costs would still far exceed revenues in 2070, and the trust fund would be dropping rapidly.

To test and calibrate our model, we ran a deterministic simulation in which we con-

strained mortality to decline to the SSA level of life expectancy in 2070, and we set the mean fertility level equal to the SSA assumed level. Our model yielded a LTAB of -2.3 percentage points, in excellent agreement with SSA's. However, our preferred forecasts follow Lee and Carter (1992) in forecasting twice as large a gain in life expectancy, leading to a mean LTAB in stochastic forecasts of -3.3 percentage points. (Comparison is complicated by differences in both the age pattern and pace of mortality decline between SSA and Lee and Carter [1992], and these differences are partially offsetting.)

SSA calculates a high-low range from 0.5 percentage points to -5.7 percentage points, for an interval width of 6.1 percentage points. Our preferred 95-percent range extends from -0.2 percentage points to -6.5 percentage points, for an interval width of 6.3 percentage points, very similar to SSA's, although the range is centered somewhat lower. Note that the SSA range reflects uncertainty in eight variables, while our range reflects uncertainty in only four.

VI. Tax Increases, Exhaustion, and Reserve Fund Constraints

According to the SSA projections, a 2.2-percentage-point increase in the OASDI tax rate should restore the system to long-term balance, under the intermediate set of assumptions. We can use our stochastic simulation to assess the likelihood that exhaustion would occur in any case. We find that, if taxes were immediately raised by 2 percentage points to 14.4 percent, 74 percent of the sample paths would still end in exhaustion by 2070. Even with a 4-percentage-point increase in the tax rate, the probability of exhaustion would still be 22 percent. To reduce the probability of exhaustion to 5 percent would require a 5-percentage-point immediate increase in taxes. These calculations are highly artificial, because in reality taxes and benefits could be adjusted in response to actual demographic and economic developments.

We find it more informative to pursue a different approach, adjusting taxes on a year-by-year basis to meet costs. The currently legislated 12.4-percent tax rate is left in place

until the reserve fund falls to 100 percent of anticipated year-ahead outflows. Thereafter, the tax rate is set each year to maintain the reserve fund at a level exactly sufficient to cover the next year's benefit payments. Figure 1 plots lines of varying probability coverage for the resulting tax rates. For example, each year, 20 percent of the simulated sample lines lie below the line labeled "20 percent" on the figure. The median line (at 50 percent) remains at 12.4 percent until 2022, rises rapidly as the baby boom retires, and then rises more slowly to 21 percent in 2070. The region with 95-percent probability coverage is bounded by the 2.5-percent and 97.5-percent lines. The lower 2.5-percent bound remains at 12.4 percent until 2044 and then rises modestly to 16 percent in 2070. The upper 97.5-percent bound rises roughly linearly from 2004, to 34 percent in 2070. There is a 2.5-percent chance that the tax rate would have to rise to 34 percent or more of earnings by 2070. A 34-percent payroll tax rate for OASDI alone would imply a very high total tax rate.

The SSA gives high, medium, and low cost rate projections of 28 percent, 19 percent, and 13 percent for 2070 (Board of Trustees, 1996 pp. 170-71), which may be compared to our projections of 34 percent, 21 percent, and 15 percent. The range widths are again similar: 15 percentage points for SSA versus 19 percentage points for the stochastic simulation. Our median forecast for 2070, 21 percent, is also fairly close to the middle forecast of the SSA, 19 percent.

VII. How Much Uncertainty Does Each Component Contribute?

It may be useful to assess the proximate sources of this considerable uncertainty in the long-term forecasts. The Board of Trustees (1996) reports a sensitivity analysis in which one of the factors at a time is varied across its high-low range, while the others are held at their middle values. Evidently, results will depend both on the sensitivity of the projection to variations and on the width of the range. The corresponding exercise for our stochastic simulation model holds all but one of the variables fixed at their mean trajectories, while allowing the remaining one to vary stochastically. This

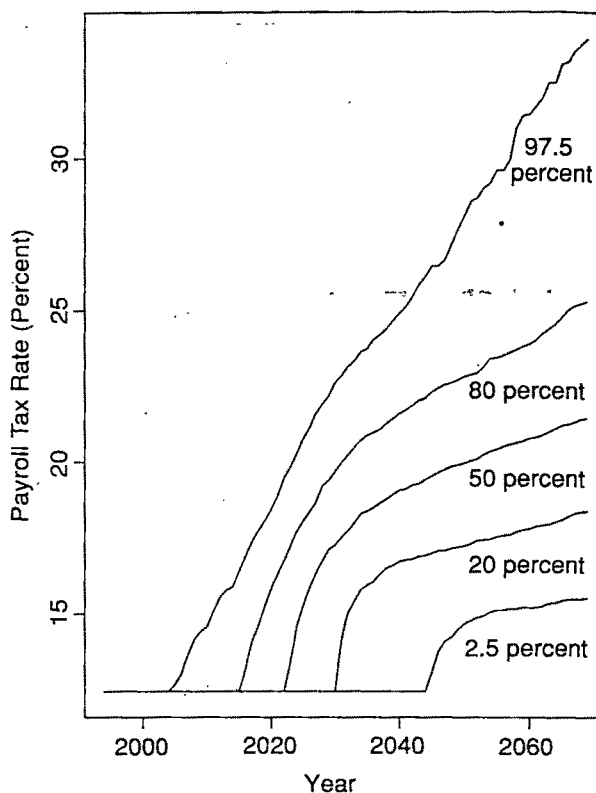


FIGURE 1. TAX RATE FOR YEAR-AHEAD BALANCE
(DISTRIBUTION BY SELECTED PROBABILITY PERCENTILES)

approach avoids arbitrary assumptions about the range of variation, which accounts for the differences described below. The standard deviation of the LTAB provides a convenient metric for comparison. We find that, through 2070, fertility contributes the greatest uncertainty, followed by productivity growth, interest, and finally mortality. While we put uncertainty about mortality last in importance, the SSA analysis (Board of Trustees, 1996 pp. 132-34) puts it first in importance. And while we put fertility first in importance, SSA puts it last (tied with the interest rate). These rankings depend strongly on the length of the period over which we compare their influence.

We can also evaluate the role of demographic uncertainty arising from fertility and mortality together. Over a 25-year horizon, demographic uncertainty generates a standard deviation only one-fifth as wide as the fully stochastic model. Over a 75-year horizon, however, demographic uncertainty alone generates a standard deviation almost two-thirds as wide as the fully stochastic model. Evidently, demographic uncertainty be-

comes more important over the long run than it is over the shorter run.

VIII. Conclusions

Many other aspects of uncertainty could be examined using these stochastic forecasts. For example, we have calculated that the cohort born during 1980–1984 has an expected rate of return from Social Security of 1.6 percent, with a 95-percent probability interval of 0.2–2.8 percent. This range corresponds to annual benefit levels differing by more than a factor of 2 for a given lifetime history of tax payments.

These simulations could also be used to evaluate the performance of policies designed to achieve various goals in the face of uncertainty: buffer the system's finances, reduce intergenerational inequities, reduce the uncertainty experienced by individual cohorts, or avoid rapid changes in taxes or benefits. Each policy could be viewed as a kind of filter, intended to attenuate certain kinds of variance in certain outcome measures. A policy rule could be applied to the set of stochastic simulations, and its success according to the target criterion could then be assessed. Is it better to wait and see, adjusting policy continuously as we gain information? Or is it better to accumulate large reserves early on, to insulate against unlikely but possible transitory demands on the system? Or should policy simply be set to deal reasonably with the mean trajectory, ignoring the uncertainty? How about keying the level of benefits to life expectancy at retirement, or tying cohort benefit levels to cohort fertility?

We are still at an early stage of formulating these stochastic forecasts, and of considering their uses. In new work, we are modeling the benefit process explicitly, including disability, rather than relying on these average age profiles. We are also developing stochastic forecasts of full long-term government budgets at

the federal and state/local levels. While much remains to be done, we believe this is a promising start.

REFERENCES

- Board of Trustees, Federal Old-Age and Survivors Insurance and Disability Insurance Trust Funds.** *1995 Annual Report of the Board of Trustees of the Federal Old-Age and Survivors Insurance and Disability Insurance Trust Funds*. Washington, DC: U.S. Government Printing Office, 1995.
- . *1996 Annual Report of the Board of Trustees of the Federal Old-Age and Survivors Insurance and Disability Insurance Trust Funds*. Washington, DC: U.S. Government Printing Office, 1996.
- Congressional Budget Office.** *The economic and budget outlook: Fiscal years 1997–2006*. Washington, DC: U.S. Government Printing Office, 1996.
- Holmer, Martin.** "Overview of SSASIM: A Long-Run Stochastic Simulation Model of Social Security." SSA Contract 95-22582, Social Security Administration, Washington, DC, 1995.
- Lee, Ronald D.** "Modeling and Forecasting the Time Series of US Fertility: Age Patterns, Range, and Ultimate Level." *International Journal of Forecasting*, August 1993, 9(2), pp. 187–202.
- Lee, Ronald D. and Carter, Lawrence.** "Modeling and Forecasting the Time Series of U.S. Mortality." *Journal of the American Statistical Association*, September 1992, 87(419), pp. 659–71.
- Lee, Ronald D. and Tuljapurkar, Shripad.** "Stochastic Population Forecasts for the U.S.: Beyond High, Medium, and Low." *Journal of the American Statistical Association*, December 1994, 89(428), pp. 1175–89.
- . "Stochastic Forecasts for Social Security," in David Wise, ed., *Frontiers in the economics of aging*. Chicago: University of Chicago Press, 1998 (forthcoming).

Demographic Analysis of Aging and Longevity

By JAMES W. VAUPEL*

The populations of most of the world's countries are growing older. This shift is creating a new demography, a demography of low fertility and long lives. The rapidly growing populations of the elderly are putting unprecedented stresses on societies, because new systems of financial support, social support, and health care have to be developed and implemented. I will focus on a particular research thrust, namely, demographic analyses of survival and longevity. I will start with a review of the remarkable improvements in survival at older ages in recent decades.

"There is one and only one cause of death at older ages. And that is old age. And nothing can be done about old age." This verbal pronouncement by Leonard Hayflick,¹ a pioneering gerontologist, captures the gist of a prevalent syndrome of beliefs. Because deaths at younger ages are now unusual in developed countries, this view implies that human life expectancy in the developed countries, and in China and many other developing countries with low mortality, is close to the limit imposed by biology. The population of older people will grow as the baby boomers age, but if this view is correct, governments need not worry that enhanced survival at older ages might accelerate the growth. Furthermore, the view that mortality at older ages is intractable leads to the conclusion that health-care resources and biomedical research should increasingly be directed toward improving "the average well being of the population" rather than extending "the average lifespan" (P. H. M. Lohman et al., 1992; see S. Jay Olshansky et al., [1990] for a subtler conclusion).

Mortality at older ages is, however, by no means intractable (John R. Wilmoth, 1997).

In fact, remarkable progress has been made since 1950 and especially since 1970 in substantially improving survival at older ages, even the most advanced ages. Despite this compelling evidence, the belief that old-age mortality is intractable remains deeply held by many people. Because of its implications for social, health, and research policy, the belief is pernicious. Because the belief is so prevalent, forecasts of the growth of the elderly population are too low, expenditures on life-saving health care for the elderly are too low, and expenditures for biomedical research on the deadly illnesses of old age are too low.

The fact is that mortality at older ages has fallen dramatically since 1950 in developed countries and most developing countries as well. For instance, over the half century from 1900 to 1950, central death rates for 85-, 90-, and 95-year-old Swedish women hovered around 0.2, 0.3, and 0.4, respectively. By 1995 these death rates had fallen below 0.1, 0.2, and 0.3. Among female octogenarians and nonagenarians in England and Wales, France, Sweden, and Japan in 1950 there were about 180 deaths per 1,000 population. By 1995 there were less than 90 deaths per 1,000 population in all four of these countries. Similar progress was achieved in most other developed countries and in many developing countries. Improvements were also made for males, although male gains have generally been smaller than female gains (Vaino Kannisto, 1994, 1996; Kannisto et al., 1994).

Another, longer-term perspective is provided in Table 1, which documents the acceleration of mortality improvements for females in the Nordic countries of Denmark, Finland, Norway, and Sweden, countries for which reliable mortality data at older ages are available well back into the 19th century. Note the rapid improvements in mortality in recent decades, especially for women in their seventies and eighties.

Table 2 displays death rates by age and time for females in the Nordic countries. The in-

* Max Planck Institute for Demographic Research, Doberaner Strasse 114, D-18057 Rostock, Germany.

¹ Talk given on 28 August 1988 at the American Association of Retired People (AARP) headquarters, Washington, DC ("The Likely Health, Longevity, and Vitality of Future Cohorts of Mid-Life and Older Persons").

TABLE 1—AVERAGE ANNUAL RATES OF IMPROVEMENT IN FEMALE MORTALITY (PERCENTAGES) FOR AGGREGATION OF DENMARK, FINLAND, NORWAY, AND SWEDEN, FOR SEXAGENARIANS, SEPTUAGENARIANS, OCTOGENARIANS, AND NONAGENARIANS, OVER SUCCESSIVE 20-YEAR PERIODS

Time period	Age category			
	Sixties	Seventies	Eighties	Nineties
1900's–1920's	0.3	0.2	0.1	0.0
1920's–1940's	0.7	0.4	0.2	0.0
1940's–1960's	1.7	1.0	0.6	0.5
1960's–1980's	1.5	2.1	1.7	1.2

Source: See Kannisto et al. (1994) for description of how average annual rates of improvement are calculated.

crease in death rates with age is striking. The decrease in death rates over time is also striking. If mortality is reduced, then the number of lives saved is proportional to the absolute decline rather than the relative decline. In the last row of Table 2, the absolute improvements in Nordic female mortality are displayed. The large absolute reductions in mortality among centenarians and nonagenarians is a remarkable achievement, at sharp variance with the view that old-age mortality is intractable.

If death rates at older ages were approaching a biological limit, then it might be expected that improvements in countries with the lowest death rates would tend to be slower than in countries with death rates further away from the irreducible minimum. There is, however, no correlation, either for males or for females, between levels of mortality and rates of mortality improvement. Furthermore, males suffer higher mortality than females, but rates of improvements for females are higher than for males (Kannisto et al., 1994).

Until recently, demographers have been wary of using U.S. mortality data at older ages because of concerns about the validity of age-reporting. New data sources now permit accurate estimation of U.S. death rates, at least up to ages in the late nineties for the white population of the United States (Bert Kestenbaum, 1992; Kenneth G. Manton and Vaupel, 1995; Laura B. Shrestha and Samuel H. Preston, 1995). It turns out that for octogenarians and nonagenarians U.S. death rates (for whites) are substantially lower than death

TABLE 2—FEMALE CENTRAL DEATH RATES (PERCENTAGES) FOR AGGREGATION OF DENMARK, FINLAND, NORWAY, AND SWEDEN, FOR SEXAGENARIANS, SEPTUAGENARIANS, OCTOGENARIANS, NONAGENARIANS, AND CENTENARIANS, IN TWO PERIODS, 1930–1949 AND 1989–1993

Period	Age category				
	Sixties	Seventies	Eighties	Nineties	100+
1930–1949	2.4	6.4	16.1	33.9	70.1
1989–1993	1.1	3.1	9.1	23.4	48.5
Change:	1.3	3.3	7.0	10.5	21.6

Source: Kannisto et al. (1994).

rates in Western Europe or Japan. In the 1980's at age 90, for instance, female death rates in Europe and Japan were almost 50-percent higher than in the Upper Midwest region of the United States (0.19 vs. 0.13) and about 20-percent higher than in the Deep South region.

This is remarkable because mortality before age 65 or 70 is substantially higher in the United States than in Western Europe and Japan. Because the very old particularly benefit from medical care and salubrious behavior, it is possible that the U.S. advantage stems from better health conditions for the elderly. The U.S. mortality advantage at older ages might also be at least in part due to the immigration of large numbers of healthy migrants into the United States in the decades before 1920. Another possibility is that conditions during childhood have lingering effects on health at advanced ages: the United States was a world leader in childhood health at the beginning of this century. In any case, the gap between the United States, on the one hand, and Western Europe and Japan, on the other, is further evidence for the plasticity of mortality at older ages.

I. Rapid Growth of the Elderly Population

I now turn to the impact of mortality reductions on the growth of the elderly population, starting with the population of centenarians. In the countries where reliable data are available on centenarians, the number of centenarians is increasing at an exceptionally rapid rate, about

8 percent per year on average. Demographers are used to population growth rates around 1 percent per year or so; an 8-percent growth rate seems more like an inflation rate. In England and Wales, an average of 74 persons per year reached age 100 between 1911 and 1920; by 1990 the number of people celebrating their 100th birthday had increased to almost 2,000, and in 1997 the number will be around 3,000 (Vaupel and Bernard Jeune, 1995). In China, Zeng Yi and I estimate that the number of centenarians is doubling every decade. In 1990 there were about 6,000 people age 100 and above in China. By the year 2000 there may be more than 12,000.

The population of centenarians is growing, in part, because of the increase in births a century ago, the sharp decline in infant and childhood mortality, and the substantial decline in mortality at ages from childhood up to age 80. Demographic analysis demonstrates, however, that by far the most important factor in the explosion of the centenarian population (two or three times more important than all the other factors combined) has been the decline in mortality after age 80 (Vaupel and Jeune, 1995).

Centenarians are still unusual, but these findings do illustrate the fact that mortality reduction can have major impacts on population growth at older ages. The growth of the population of female octogenarians in England and Wales provides another telling example. The remaining life expectancy of 80-year-old females in England and Wales around 1950 was approximately six years. Currently the corresponding figure is about nine years, some 50-percent higher. As a result, the population of female octogenarians in England and Wales is roughly half again as big as it would have been if mortality after age 80 had remained at 1950 levels. Putting this in terms of population counts, more than a half million females aged 80+ are alive today in England and Wales who would have been dead if mortality after age 80 had not been reduced.

Table 3 provides information about the size of the older population of various countries, from age 60 and up, for both sexes combined. Estimates are also given for the size of these populations in 2025. The size of the older population shows substantial increases not only

TABLE 3—PROPORTION OF POPULATION ABOVE AGE 60 (PERCENTAGE) AND POPULATION ABOVE AGE 60 (IN MILLIONS) FOR SELECTED COUNTRIES IN 1996 AND PROJECTED FOR 2025

Country	Age 60+ (percentage)		Age 60+ (millions)	
	1996	2025	1996	2025
Italy	22	33	13	18
Japan	21	33	26	40
Germany	21	32	17	28
France	20	30	12	18
United Kingdom	21	29	12	17
United States	17	25	44	83
China	9	20	115	290
Brazil	7	16	11	31
Mexico	7	13	6	18
India	7	12	62	165
South Africa	7	10	3	6
Egypt	6	10	4	10

Source: U.S. Bureau of the Census (1997).

in Europe, Japan, and the United States, but China, India, and other developing countries as well.

II. Variation in Lifespan

The multiplication of the population of older people heightens interest in a fundamental question: why do some people die at 60, others at 80, and a few at 100? Why are the odds of dying at 80 rather than 60 increasing and the chance of surviving to 100 rapidly increasing (albeit from a very low level)? How important are genetic versus environmental, behavioral, and medical factors in determining how long an individual will live?

Studies of twins and other kinds of related individuals suggest that about 25 percent of the variation in adult lifespans appears to be attributable to genetic variation among individuals (Matt McGue et al., 1993; Anne Maria Herskind et al., 1996). Some research in progress by two of my colleagues (Anatoli Yashin and Ivan Iachine) suggests that an additional 25 percent may be attributable to nongenetic characteristics that are more or less fixed by the time a person is 30 or so: characteristics such as educational achievement, socioeconomic status, mother's and father's age at a person's birth, etc. Research on the relative

importance for longevity of various candidate genes and nongenetic fixed attributes is, however, still at an early stage of development.

David J. P. Barker's (1992, 1995) "fetal-origins hypothesis" suggests that nourishment in utero and during infancy programs the development of risk factors for several important diseases of middle and old age. Other researchers have also concluded that nutrition and infections early in life have major effects on adult mortality (W. Kermack et al., 1934; Irma T. Elo and Preston, 1992; Fogel, 1993). To the extent that this is true, longevity may be determined by conditions in childhood and perhaps before birth. There is, however, conflicting evidence that suggests that current conditions (i.e., at older ages) may be much more important than conditions early in life. Kannisto (1994, 1996) finds period effects to be considerably more significant than cohort effects on mortality after age 80. Kaare Christensen et al. (1995) find that, from age 6 up to the oldest ages, twins (who tend to be born prematurely and at low birth weight) suffer the same age-specific death rates as singletons; and Kannisto et al. (1997) find no increased mortality in later life for cohorts born during the Finnish famine of 1866–1868. Pinning down the nature and magnitude of possible lingering effects of early-life conditions on survival at advanced ages is an important research priority.

III. Trajectories of Mortality at Advanced Ages

Further insights into the determinants and plasticity of longevity can be gleaned by analyzing the trajectories of age-specific death rates at advanced ages, both for humans and for various nonhuman species (Vaupel, 1997). Benjamin Gompertz (1825) proposed that the force of mortality increased exponentially with age for humans, at least as a serviceable approximation over the range of adult ages for which he had data. Various subsequent researchers, especially in biology and gerontology, have viewed Gompertz's observation as a law that describes the process of senescence in almost all multicellular animals at all ages after the onset of reproduction. As a rough approximation at younger adult ages, Gompertz's exponential formula does capture

the rise in mortality in a great variety of species (Caleb E. Finch, 1990). Human mortality, however, does not increase exponentially after age 80. Mortality decelerates, rising perhaps to a maximum or ceiling around age 110 (A. Roger Thatcher et al., 1998). Whether mortality is slowly increasing, level, slowly decreasing, or rapidly decreasing after age 110 is uncertain.

Humans are animals. Almost all animals show signs of aging, and for almost all animals death rates tend to rise after the age of maturity (Finch, 1990). Even researchers who are only interested in people may benefit from biological insights from studies of other species, because these insights may cast light on the biology of humans. I will give just one example. The largest nonhuman population followed to natural death consisted of 1.2 million medflies studied in a laboratory near Tapachula, Mexico. These flies were held in cages, each holding several thousand flies. As reported by James R. Carey et al. (1992), the trajectory of mortality rose, peaked, and then fell to a low level around which it hovers until the last fly died at an age of 171 days (compared with an average life span in the experiment of 21 days).

Why does mortality decelerate? One reason is that all populations are heterogeneous. Some individuals are frailer than others, and the frail tend to die first. This creates a fundamental problem—indeed, it seems to me the fundamental problem—for demographic analyses in general and for analyses of age-trajectories of mortality in particular. The individuals alive at older ages are systematically different from the individuals alive at younger ages. The age-trajectory of mortality reflects both the underlying age-trajectories of mortality for individuals in the population and the effects of compositional change as the frailer individuals drop out of the population (e.g., Vaupel et al., 1979; Vaupel and Anatoli I. Yashin, 1985; Vaupel and Carey, 1993).

Living organisms are complex systems. Reliability engineers and systems analysts have learned a great deal about the failure of complex systems. Automobiles are popular pieces of complicated equipment. They are sufficiently standardized that it is meaningfully possible to count their numbers on an

age-specific basis. Then age-specific death rates can be calculated. It turns out that automobile mortality rises steeply at younger ages but levels off around age 10 or 12 (Vaupel and Cindy R. Owens, 1997). The question arises: is mortality a property of living organisms or a property of complicated systems? When it comes to death, how do people and flies differ from Toyotas? In particular, is the deceleration and leveling off of mortality a fairly general property of complicated systems? Better understanding of these questions may lead to new insights into aging and survival.

IV. Conclusion

Over the past half century, and especially in the most recent decades, remarkable improvements have been achieved in survival at older ages, especially at the highest ages. This progress has accelerated the growth of the population of older people and has advanced the frontier of human survival substantially beyond the extremes of longevity attained in pre-industrial times. The widely held position that mortality at older ages is intractable is untenable. However, little is yet known about why mortality among the oldest-old has been so plastic since 1950. There is considerable (but still inadequate) knowledge of why some people die in infancy or childhood and why some people die prematurely at adult ages before age 60 or 70. Much less is known about why some people survive to age 80, others to age 90, and a few to age 100. The little that is known has largely been learned within the past few years, and new findings (especially concerning genetic factors) are emerging at a rapid rate. A key finding is that mortality decelerates at advanced ages not only for humans, but for other biological species and for automobiles as well. The deceleration results from some mix of genetic, environmental, behavioral, bio-reliability, and heterogeneity forces and constraints, but the mix is not well understood.

REFERENCES

- Barker, David J. P. "Fetal and Infant Origins of Adult Disease." *British Medical Journal*, 1992, 301(6761), pp. 1111-14.
- . "Fetal Origins of Coronary Heart Disease." *British Medical Journal*, 1995, 311(6998), pp. 171-74.
- Carey, James R.; Liedo, Pablo; Orozco, Dina and Vaupel, James W. "Slowing of Mortality Rates at Older Ages in Large Medfly Cohorts." *Science*, 16 October 1992, 258(5081), pp. 457-61.
- Christensen, Kaare; Vaupel, James W.; Holm, Niels V. and Yashin, Anatoli I. "Mortality among Twins after Age 6: Fetal Origins Hypothesis versus Twin Method." *British Medical Journal*, 18 February 1995, 310(6977), pp. 432-36.
- Elo, Irma T. and Preston, Samuel H. "Effects of Early-Life Condition on Adult Mortality: A Review." *Population Index*, Summer 1992, 58(2), pp. 186-222.
- Finch, Caleb E. *Longevity, senescence, and the genome*. Chicago: University of Chicago Press, 1990.
- Fogel, Robert W. *Economic growth, population theory, and physiology: The bearing of long-term processes on the making of economic policy*. Stockholm, Sweden: Nobel Foundation, 1993.
- Gompertz, Benjamin. "On the Nature of the Function Expressive of the Law of Human Mortality." *Philosophical Transactions*, 1825, 27, pp. 510-19; reprinted in David Smith and Nathan Keyfitz, eds., *Mathematical demography: Selected papers*. New York: Springer-Verlag, 1977, pp. 279-82.
- Herskind, Anne Maria; McGue, Matt; Holm, Niels V.; Soerensen, Thorkild, I. A. and Vaupel, James W. "The Heritability of Human Longevity." *Human Genetics*, March 1996, 97(3), pp. 319-23.
- Kannisto, Vaino. *Development of oldest-old mortality, 1950-1990*. Odense, Denmark: Odense University Press, 1994.
- . *The advancing frontier of survival: Life tables for old age*. Odense, Denmark: Odense University Press, 1996.
- Kannisto, Vaino; Christensen, Kaare and Vaupel, James W. "No Increased Mortality in Later Life for Cohorts Born During Famine." *American Journal of Epidemiology*, June 1997, 145(11), pp. 987-94.
- Kannisto, Vaino; Lauritsen, Jens; Thatcher, A. Roger and Vaupel, James W. "Reductions in Mortality at Advanced Ages." *Population*

- and Development Review*, December 1994, 20(4), pp. 793–810.
- Kermack, W.; McKendrick, A. and McKinlay, P. "Death-Rates in Great Britain and Sweden: Some General Regularities and Their Significance." *Lancet*, 31 March 1934, 226 (5770), pp. 698–703.
- Kestenbaum, Bert. "A Description of the Extreme Aged Population Based on Improved Medicare Enrollment Data." *Demography*, November 1992, 29(4), pp. 565–80.
- Lohman, P. H. M.; Sankaranarayanan, K. and Ashby, J. "Choosing the Limits to Life." *Nature*, 21 May 1992, 357(6375), pp. 185–86.
- Manton, Kenneth G. and Vaupel, James W. "Survival after the Age of 80 in the United States, Sweden, France, England, and Japan." *New England Journal of Medicine*, 2 November 1995, 333(18), pp. 1232–35.
- McGue, Matt; Vaupel, James W.; Holm, Niels and Harvald, Bert. "Longevity Is Moderately Heritable in a Sample of Danish Twins Born 1870–1880." *Journal of Gerontology*, November 1993, 48(6), pp. B237–44.
- Olshansky, S. Jay; Carnes, Bruce A. and Cassel, Christine. "In Search of Methuselah: Estimating the Upper Limits of Human Longevity." *Science*, 2 November 1990, 250(4981), pp. 634–40.
- Shrestha, Laura B. and Preston, Samuel H. "Consistency of Census and Vital Registration Data on Older Americans: 1970–1990." *Survey Methodology*, 1995, 21(2), pp. 167–77.
- Thatcher, A. Roger; Kannisto, Vaino and Vaupel, James W. *The force of mortality at ages 80 to 120*. Odense, Denmark: Odense University Press, 1998.
- U.S. Bureau of Census. *Global aging into the 21st century*. Washington, DC: U.S. Bureau of the Census, 1997.
- Vaupel, James W. "Trajectories of Mortality at Advanced Ages," in Kenneth W. Wachter and Caleb E. Finch, eds., *Between Zeus and the salmon: The biodemography of longevity*. Washington, DC: National Academy Press, 1997, pp. 17–37.
- Vaupel, James W. and Carey, James R. "Compositional Interpretations of Medfly Mortality." *Science*, 11 June 1993, 260(5114), pp. 1666–67.
- Vaupel, James W. and Jeune, Bernard. "The Emergence and Proliferation of Centenarians," in Bernard Jeune and James W. Vaupel, eds., *Exceptional longevity: From prehistory to the present*. Odense, Denmark: Odense University Press, 1995, pp. 109–16.
- Vaupel, James W.; Manton, Kenneth G. and Stallard, Eric. "The Impact of Heterogeneity in Individual Frailty on the Dynamics of Mortality." *Demography*, August 1979, 16(3), pp. 439–54.
- Vaupel, James W. and Owens, Cindy R. "Automobile Demography," Unpublished manuscript presented at the 23rd IUSSP General Population Conference, Beijing, China, 11–17 October 1997.
- Vaupel, James W. and Yashin, Anatoli I. "Heterogeneity's Ruses: Some Surprising Effects of Selection on Population Dynamics." *American Statistician*, July 1985, 39(3), pp. 176–85.
- Wilmoth, John R. "In Search of Limits: What Do Demographic Trends Suggest about the Future of Human Longevity," in Kenneth W. Wachter and Caleb E. Finch, eds., *Between Zeus and the salmon: The biodemography of longevity*. Washington, DC: National Academy Press, 1997, pp. 38–64.

Aging and Inequality in Income and Health

By ANGUS S. DEATON AND CHRISTINA H. PAXSON*

In our previous work, Deaton and Paxson (1994, 1997), we showed that, in a large group of countries, inequality in consumption increases with age within cohorts of individuals. This finding was motivated by a well-known feature of standard autarkic intertemporal choice models, that under appropriate assumptions consumption follows a martingale (see Robert E. Hall, 1978). The theory implies that within-cohort consumption inequality should rise over time as cohorts age, provided that shocks to consumption are not perfectly correlated across individuals. The same should be true of income, at least up to the date of retirement, and of earnings, if employers pay workers their expected marginal product (see Henry S. Farber and Robert Gibbons, 1996).

More recently we have examined whether inequality in health status also increases with age, and how the joint distribution of health and income evolve over the life cycle. It is plausible that health shocks have both permanent and transitory components. The presence of the former implies that health status will be nonstationary so that, provided health shocks are not perfectly correlated across individuals, the dispersion of health status will grow with age. This view of health status as a nonstationary random variable is consistent with stress models in which poor health is the result of "the piling up of adverse life experiences" (Carol D. Ryff and Burton Singer, 1997 p. 90).

Health status, along with income and consumption, is an important determinant of welfare, so that our interest in health inequality stems from a more general interest in the distribution of welfare. Furthermore, health is not

independent of economic status. There is a well-documented but poorly understood "gradient" linking socioeconomic status to a wide range of health outcomes (see Nancy E. Adler et al. [1994] and Sally MacIntyre [1997] for reviews). The gradient has both a life-cycle and a temporal component; differences in mortality across socioeconomic groups are widest in late middle age Evelyn M. Kitigawa and Philip M. Hauser, 1973; Harriet Orcutt Duleep, 1995; Irma T. Elo and Samuel H. Preston, 1996) and are increasing over time (Jacob J. Feldman et al., 1989; G. Pappas et al., 1993; Preston and Elo, 1995).

In our earlier work, we used data from the National Health Interview Survey (NHIS) to examine life-cycle patterns in health status and in the joint distribution of health status and income (Deaton and Paxson, 1998). In this paper we summarize and extend those results and provide new evidence from the Panel Study of Income Dynamics (PSID). Both surveys contain a measure of household income and collect information on an ordinal measure of self-reported health status (SRHS) that ranges from 1 (excellent) to 5 (poor).

I. Measurement Issues

The measurement of health inequality raises two important issues. The first is the difficulty of identifying a measure of health status that is useful over the complete adult life cycle. For example, measures of the inability to complete "activities of daily living" (ADL's), such as dressing or bathing, have been fruitfully used to assess the health of the elderly. However, these measures do not adequately capture health differences among younger people. Self-reported "days of illness" or "doctor visits" are themselves conditioned by socioeconomic status and sometimes show perverse correlations with income, with better-off people apparently perceiving and treating their illnesses more seriously. The properties of the measure of self-reported health status used in

* Research Program in Development Studies, Princeton University, Princeton NJ 08544. The research was supported by the National Institute of Aging through grants P01-AG05842 and R01-AG11957. We thank our discussant, James Poterba, for useful comments.

this paper have been studied extensively. First, it predicts subsequent mortality. A large number of studies that use data from a variety of countries indicate that reports of poor health are significantly related to subsequent mortality (see E. L. Idler and S. V. Kasl [1995] for summary of this research). The correlation between SRHS and subsequent mortality remains strong even after controlling for objective measures of health status (obtained from doctors' examinations) and life-style factors such as smoking. This fact has led some to argue that SRHS is itself an independent determinant of longevity: individuals with healthier self-images live longer. An alternative to these psychosocial explanations is that individuals have information about their health that is unobserved by others, including physicians. For our purposes, it is the raw correlations between self-rated health and mortality that are of interest, since we want to identify a variable that can serve as a single summary measure of health status. Other research has found that those with low SRHS are more likely to develop problems with ADL's (Idler and Kasl, 1995) and miss more work due to illness (M. Marmot et al., 1995).

Once a measure of health is identified, the second issue is how to measure inequality in health status. Although it is straightforward to compute measures of dispersion in SRHS, it is not clear how we should judge such measures in terms of social welfare. Consider, for example, the familiar result that, if a distribution F_1 (second-order) stochastically dominates a distribution F_2 , then F_1 will result in higher social welfare, when social welfare is represented as the integral over the population of a monotone increasing and concave function of the variable in question. Although we are used to the assumption that social welfare is increasing and concave in income or consumption, it is much less clear why it should be increasing and concave in an ordinal self-reported measure of health status. Nevertheless, the literature on SRHS provides some support for the idea that changes in SRHS have a larger effect on mortality when SRHS is "poor" than when it is "excellent." If so, a mean-preserving spread in SRHS will lower average life expectancy and will lower the expected value of any function that is concave in

life expectancy, for example, one that prefers a decrease in infant mortality to an increase in longevity at older ages. Of course, to focus solely on life expectancy ignores the quality of life. SRHS may well give some indication of quality as well as the likely length of life, so that changes in the distribution of SRHS could still have adverse welfare consequences even in the absence of a relationship between SRHS and mortality. We also note that much of the literature on health inequality is not concerned with inequality in years lived, but with the inequalities in health outcomes across socioeconomic groups. That these are quite different has been emphasized by Richard G. Wilkinson (1986) who points out that, over the 20th century in Britain, socioeconomic differences in mortality have increased while the inequality in years lived has decreased, essentially because of the decline in infant mortality.

II. Evidence on Life-Cycle Patterns in the Distribution of Health and Income

The NHIS is an annual survey of approximately 50,000 adults (plus children) that collects information on health, illnesses, doctor visits, spells of hospitalization, and basic socioeconomic characteristics. We use data on all adults between the ages of 20 and 70, inclusive, interviewed from 1983 through 1994. The survey provides sampling weights, which we use, so that the results should be representative of the whole U.S. population. The PSID is a panel survey of households that has been in existence since 1968, and since 1984 it has collected information on the self-reported health status of household heads and their spouses. We use a sample of 3,435 men and 4,561 women who were either heads of households or their spouses in all years between 1984 and 1992. Unlike the NHIS, this is not a nationally representative sample of all adults, both because it is only households heads and spouses, and because the PSID oversampled poor households in 1968. Given these circumstances, we did not use sampling weights with the PSID.

Although the NHIS has much larger sample sizes and more extensive health information than the PSID, it has far less

information on income. The measure of family income in the NHIS is bracketed and is top-coded at \$50,000 in nominal dollars. The brackets are narrow and are not a major concern, but such serious top-coding cannot be ignored in computing measures of dispersion. In the NHIS results that follow, we have used the Tobit procedure described in Deaton and Paxson (1998), but one reason for extending our work to the PSID is to attempt to reproduce our results with much higher-quality data on income.

Our approach is to track the moments and co-moments of SRHS and family income over time for individuals from the same birth cohort. The NHIS is large enough for each cohort to be defined by the exact year of birth; for the PSID we define cohorts using nonoverlapping five-year birth intervals. Cohorts are identified by their age (or, for the PSID, the midpoint of the age range) in the earliest year of observation; 1983 for the NHIS and 1984 for the PSID. There are 62 cohorts for the NHIS, and nine for the PSID. It should be kept in mind that the PSID is used to construct "true" cohorts: we actually follow the same individuals over time as they age. With the NHIS, we track randomly selected representatives from the (still-living) populations of people born in the same year. These populations are not fixed, because some group members die each year. The evolution of inequality in health and income with age will reflect both changes in inequality within a fixed group and the effects caused by selection of some members, through death, out of the group.

We first compute moments (mean and variances) and co-moments of health status and income, for each cohort in each year, and for men and women separately. These become the "raw data" for our analysis, and much can be learned by looking at graphs of these data. Figures 1 and 2 show the cohort plots for males and females from the PSID; the same information for the NHIS is in figure 4 of Deaton and Paxson (1998). The figures show the age-profiles for the mean of SRHS (top panels), the variance of health status (middle panels), and the correlation between health status and income (bottom panels). Each line on the graphs shows the experience of a single cohort over time.

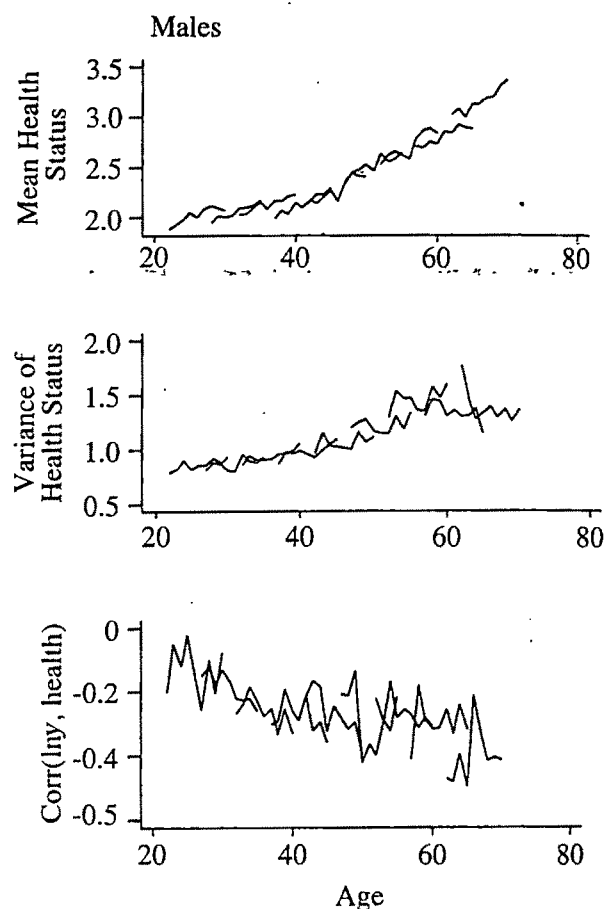


FIGURE 1. HEALTH STATUS, THE VARIANCE OF HEALTH STATUS, AND THE CORRELATION BETWEEN HEALTH AND INCOME FOR COHORTS IN THE PSID, MALES

Note first that, as expected, average health status deteriorates with age for both men and women, although women report worse health than men at younger ages. That SRHS worsens with age is perhaps to be expected, but it implies that when people report their health status, they do not completely "norm" their answers with respect to the experience of those at the same age. The patterns of SRHS with age in the NHIS are similar, except health is better on average at all ages for men and women, which is perhaps not surprising given the oversampling of poor households by the PSID. Second, inequality in health increases with age, and the results for the PSID in the middle panels of Figures 1 and 2 are consistent with the evidence from the NHIS. Although we do not show it, in both the PSID and the NHIS the dispersion in the joint distribution of income and health status rises with age up to the age of retirement and then levels off.

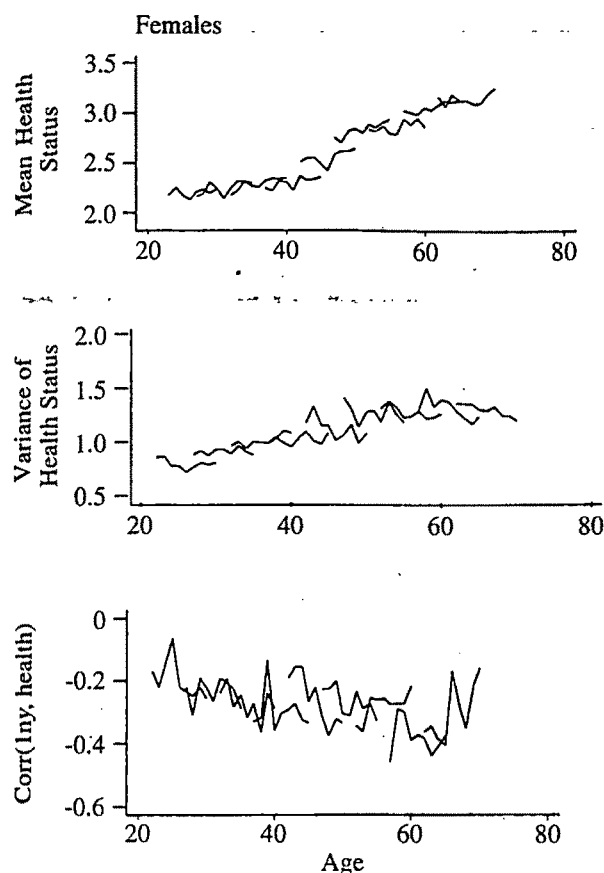


FIGURE 2. HEALTH STATUS, THE VARIANCE OF HEALTH STATUS, AND THE CORRELATION BETWEEN HEALTH AND INCOME FOR COHORTS IN THE PSID, FEMALES

Third, the bottom panels show a consistently negative correlation between health status (measured negatively) and the logarithm of family income, so that the gradient between mortality and income extends to SRHS. Moreover, and again in line with the literature, the correlation varies with age; it is small among those in their early twenties but becomes steadily larger (in absolute value), reaching a peak value of around -0.4 between ages 50 and 60. The small sample sizes in the PSID (relative to the NHIS) yield only noisy measures of this correlation; to clarify the results, and to facilitate comparisons between the two data sets, we regressed the correlations between health and income on a set of age and cohort dummy variables. The cohort dummies account for the fact that the correlation between the two variables (the gradient) may differ across groups born in different years, while the age effects capture the life-cycle pro-

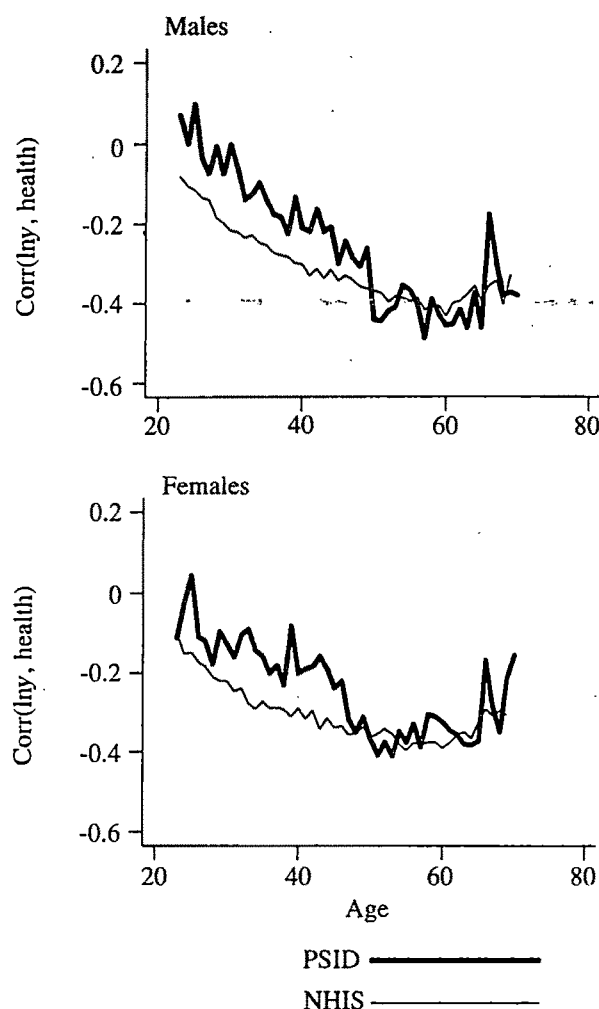


FIGURE 3. ESTIMATED AGE EFFECTS IN THE CORRELATION BETWEEN SRHS AND LOG INCOME (THE GRADIENT)

file of the gradient, the shape of which, by assumption, is held fixed across cohorts. Figure 3 shows the estimated age effects in both data sets.

The two data sets yield very similar patterns in the age profile of the correlation between income and health. For both men and women, the correlation between SRHS and income weakens after age 60, as SRHS deteriorates in general. But this is not simply a matter of the elderly having uniformly poor health status. As the top panels of Figures 1 and 2 show, health status deteriorates with age, but the middle panels do not show a collapse in the variance after age 60; instead, the fact is that, at older ages, differences in SRHS are less well-predicted by income.

There are several possible interpretations of these results, none of which necessarily excludes any other. One is that labor supply and earnings ability are adversely and cumulatively affected by health shocks, so that poor health and low income are increasingly correlated with age. The correlation may weaken in old age, since health shocks received after retirement will not affect pensions and Social Security (although they could affect asset income if sick people run down assets to pay for care). There is also undoubtedly some causality running from income to health. Poorer people are more prone to lifestyles with enhanced risk factors (e.g., obesity or cigarette smoking), have less access to health care, including preventative health care, and live and work in less healthy environments. There is also a literature documenting the adverse health consequences of unemployment. The provision of Medicare at older ages could reduce the correlation by making one determinant of health, medical care, available to everyone. Sorting out the respective contributions of these various mechanisms remains an important task for future research.

Perhaps even more important than life-cycle patterns is the question of changes over time in the relationship between income and health status. It is difficult to discern cohort effects, represented by upward or downward shifts in the traces for different cohorts, from a visual inspection of Figures 1 and 2. However, a more systematic approach shows that there are significant differences across cohorts. We first examined the cohort effects from regressions of each of the variables (mean health, the variance of health, and the correlation of health and income) on sets of age and cohort dummy variables. This was done separately for men and women, and for the PSID and the NHIS. The estimated cohort effects are jointly significant at the 5-percent level or better for each of the variables and samples. To summarize the size and sign of these cohort effects, we then regressed each of the variables on a complete set of age dummies and a linear cohort trend. The results are as follows. First, for females from both data sets and males in the NHIS, average health status has improved over time across cohorts. The effect is small: approximately 0.004 units

per year of birth. However, a visual inspection of the cohort effects indicates that they are *not* linear. There has been no improvement, and possibly some deterioration, in health status across cohorts born after 1945, and there were larger improvements across those born before 1945. The results for the males in the PSID are at odds with the other groups. The estimate of the cohort trend indicates that more recently born groups are significantly less healthy, by about 0.009 units per year of birth. This is largely due to declines in reported health, controlling for age, of the youngest four cohorts. These declines, which can be seen in the raw data graphed in the top panels of Figures 1 and 2, warrant further analysis. Second, for all of the samples, younger cohorts have a lower variance of health status.

Third, the results provide some support for the findings cited above that the gradient between income and health is becoming stronger over time. The coefficient on the cohort trend ranges from 0.001 (for females in the NHIS) to 0.003 per year (for males in the PSID), a positive sign indicating that for more recently born cohorts there is a larger correlation between income and health. For example, the actual correlation between the logarithm of income and SRHS is -0.40 for PSID males born between 1940 and 1944, when they reached the age of around 50 in 1992. Our results imply that the correlation for the cohort born ten years later, between 1950 and 1954, will equal -0.43 when this group reaches the age of 50.

REFERENCES

- Adler, N. E.; Boyce, T.; Cohen, S.; Folkman, S.; Kahn, R. L. and Syme, S. L. "Socioeconomic Status and Health: The Challenge of the Gradient." *American Psychologist*, January 1994, 49(1), pp. 15-24.
- Deaton, Angus S. and Paxson, Christina H. "Intertemporal Choice and Inequality." *Journal of Political Economy*, June 1994, 102(3), pp. 437-67.
- . "The Effects of Economic and Population Growth on National Saving and Inequality." *Demography*, February 1997, 34(1), pp. 97-114.

- _____. "Health, Income, and Inequality over the Life-Cycle," in David Wise, ed., *The economics of aging*, Vol. 7 (preliminary title). Chicago: University of Chicago Press, 1998 (forthcoming).
- Duleep, Harriet Orcutt. "Mortality and Income Inequality among Economically Developed Countries." *Social Security Bulletin*, Summer 1995, 58(2), pp. 34-50.
- Elo, Irma T. and Preston, Samuel H. "Educational Differentials in Mortality: United States, 1979-85." *Social Science and Medicine*, January 1996, 42(1), pp. 47-57.
- Farber, Henry S. and Gibbons, Robert. "Learning and Wage Dynamics." *Quarterly Journal of Economics*, November 1996, 111(4), pp. 1007-47.
- Feldman, Jacob J.; Makuc, Diane M.; Kleinman, Joel C. and Cornoni-Huntley, Joan. "National Trends in Educational Differentials in Mortality." *American Journal of Epidemiology*, May 1989, 129(5), pp. 919-33.
- Hall, Robert E. "Stochastic Implications of the Life Cycle-Permanent Income Hypothesis: Theory and Evidence." *Journal of Political Economy*, December 1978, 86(6), pp. 971-87.
- Idler, E. L. and Kasl, S. V. "Self-Ratings of Health: Do They Also Predict Change in Functional Ability?" *Journal of Gerontology*, November 1995, 50(6), pp. S344-53.
- Kitigawa, Evelyn M. and Hauser, Philip M. *Differential mortality in the United States: A study in socioeconomic epidemiology*. Cambridge, MA: Harvard University Press, 1973.
- Marmot, M.; Feeney, A.; Shipley, M.; North, F. and Syme, S. L. "Sickness Absence as a Measure of Health Status and Functioning: From the UK Whitehall II Study." *Journal of Epidemiology and Community Health*, April 1995, 49(2), pp. 124-30.
- MacIntyre, Sally. "The Black Report and Beyond: What are the Issues?" *Social Science and Medicine*, March 1997, 44(6), pp. 723-45.
- Pappas, G.; Queen, S.; Hadden, W. and Fisher, G. "The Increasing Disparity in Mortality between Socioeconomic Groups in the United States, 1960 and 1986." *New England Journal of Medicine*, July 1993, 329(2), pp. 103-15.
- Preston, Samuel H. and Elo, Irma T. "Are Educational Differentials in Adult Mortality Increasing in the United States?" *Journal of Aging and Health*, November 1995, 7(4), pp. 476-96.
- Ryff, Carol D. and Singer, Burton. "Racial and Ethnic Inequalities in Health: Environmental, Psychosocial and Physiological Pathways," in B. Devlin, S. E. Fienberg, D. P. Resnick, and K. Roeder, eds., *Intelligence, genes, and success: Scientists respond to the Bell Curve*. New York: Springer-Verlag, 1997, pp. 89-122.
- Wilkinson, Richard G., ed. *Class and health: Research and longitudinal data*. London: Tavistock, 1986.

INTERGENERATIONAL RELATIONS[†]

Generations and the Distribution of Economic Well-Being: A Cross-National View

By TIMOTHY M. SMEEDING AND DENNIS H. SULLIVAN*

This brief paper explores differences in economic well-being across cohorts of the population in four modern nations (Canada, Sweden, the United Kingdom, and the United States). It focuses on relative incomes within countries, poverty rates, and social expenditures by age group over the 1974–1994 period. Cross-national patterns of level (data from the most recent year) and trend are both explored.

This work follows in the long tradition of Richard Easterlin to examine patterns of economic change over the life cycle from a cross-national perspective. Ultimately, economists are primarily interested in the ways in which individuals respond to economic conditions in terms of choice of housing arrangements, marriage, childbearing, and work status via decisions such as schooling and retirement (Easterlin, 1987). And our work will eventually explore each of these arenas from a cross-national perspective. Therefore, this paper should be seen as an appetizer for a wider range of in-depth explorations to follow.

I. Data and Technical Details

The data for this paper come from the Luxembourg Income Study (LIS) data base, a collection of almost 100 household surveys in 27

countries spanning the 1970–1995 period. LIS has been widely used over the past decade to examine relative economic well-being, poverty, and inequality across a large number of nations (Peter Gottschalk and Smeeding, 1997). Here we select four nations because of their similarities and differences with respect to size, language, social-security institutions, and pattern of social expenditure. The United States and the United Kingdom have experienced rapid secular increases in inequality and in relative poverty over this period. Inequality has increased less in Canada and Sweden, and poverty levels are generally lower in these nations as well.

Our unit of account is the household, and our main measure of well-being is household size-adjusted disposable income (ADPI), where after-tax and transfer disposable income (DPI) is adjusted for differences in household size (S) based on a logarithmic equivalence scale:

$$ADPI = DPI/S^e$$

where $e = 0.5$. The unit of analysis is the household for income and social transfers, but persons (within households) for poverty measurement. Households were grouped by age of household head into five-year age ranges.

We employ three measures to examine the relative situation of age groups:

- (i) Relative median income (“income”), which is the ratio of the ADPI of the household containing the person with the median income within an age group to the ADPI of the household containing the median-income person among all households.
- (ii) Relative poverty (“poverty”), which is

[†] *Discussants:* Jere R. Behrman, University of Pennsylvania; Marilyn Moon, Urban Institute.

* The authors thank Richard Easterlin, Jere Behrman, and Marilyn Moon for comments, and the NSF and NIA for financial support. A longer version of this paper with complete numerical data is available under the same title as LIS Working Paper No. 173 on the LIS home page at:

<http://lissy.ceps.lu/wpapers.htm>

the fraction of the persons in an age group with incomes less than half of the ADPI of the household containing the median-income person among all households. Thus, this measure shows the percentage of persons living with a head of a given age where the household income is less than half the median for all households.

- (iii) Social transfers ("transfer"), which are the ratios of total cash social expenditures received by an age group to the sum of pretax, pre-transfer market incomes received by that age group. Market income is top- and bottom-coded to ensure consistency over time and across nations.

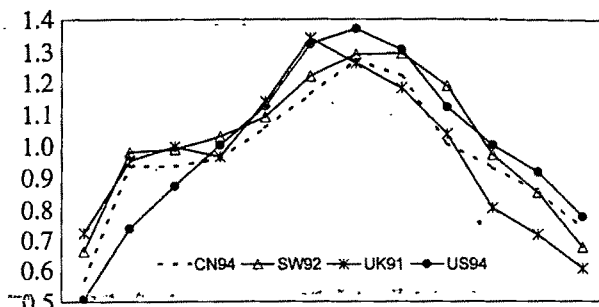
The basic data are presented in two figures. Trend data, which are not shown, are available from the authors or from the LIS website (see * footnote on preceding page).

II. Relative Income and Poverty: Levels and Trends

Patterns of (relative median) income and poverty are examined across countries and over time within countries. Each country's age-income profile displays the familiar inverted U shape (Fig. 1A), though the peaks occur at different points across the nations: earlier (ages 45–49) in the United Kingdom, and later (ages 55–59) in Sweden, with Canada and the United States peaking in between (ages 50–54). The U.S. profile is also more peaked than those of the other nations, particularly owing to the relatively lower incomes of younger households in the United States. The second-most peaked profile is that of the United Kingdom, reflecting the rapid decline in average incomes among its aged. The profile declines least rapidly with age in the United States.

Overall levels of inequality differ markedly across these nations, and these differences are reflected in their poverty rates. The United States has the highest relative poverty rates, and Sweden the lowest, among the age groups shown here up to age 65 (Fig. 1B). Beyond age 65, elders in the United Kingdom have higher poverty rates than do elders in the United States; and beyond age 70, Canada dis-

A. Relative Median Income



B. Relative Poverty

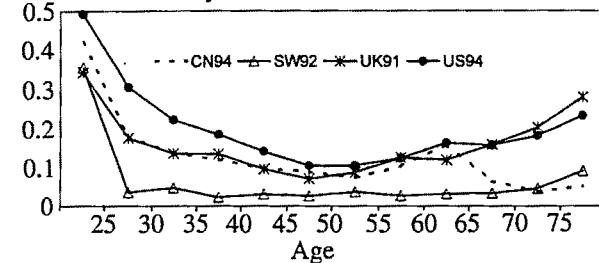


FIGURE 1. RELATIVE MEDIAN INCOME AND RELATIVE POVERTY BY AGE OF HOUSEHOLD HEAD IN THE EARLY 1990'S FOR FOUR NATIONS

Notes: Relative income is the adjusted disposable income of the household containing the median-income person in the age group relative to the adjusted disposable income of the household containing the median-income person in the nation. Relative poverty is the poverty rate in the cohort, where a household's members are "poor" if the adjusted disposable income of the household is less than half of the adjusted disposable income of the household containing the median-income person in the nation.

plays a lower elder poverty rate than does Sweden. The overall poverty rates are 6.7 percent in Sweden, 11.6 percent in Canada, 14.6 percent in the United Kingdom, and 19.1 percent in the United States.

The age pattern of poverty that emerges is clear. In general, younger households (under age 30) are doing worse than are elders (over age 65), and both the young and the old have higher poverty rates in the United States than in other nations. Middle-aged families, aged 40–54, who generally represent the baby boomers in most of these nations, tend to have the highest incomes and lowest poverty rates in every nation.

Trends within each nation differ. There are some large movements in both incomes and poverty over time. The relative position of young households in the United States has worsened considerably since 1979, producing

a much steeper age-income profile through the working ages. This pattern is consistent with other analyses of this age group using other data sources (e.g., David Card and Thomas Lemieux, 1997; Greg Duncan et al., 1996). In contrast, the aged have greatly improved their relative income position, with the largest differences occurring at older ages. In general, changes in poverty rates mirror this worsening for the young and the improvement in economic status among the aged.

The evolution of the age-income profile in the United Kingdom has been similar to that in the United States, but much less dramatic. Although the relative median position of younger households has not changed greatly, poverty is generally higher across the entire working-age population than it was in the 1970's, particularly among the young. This result is due, in large part, to a general rise in British inequality over this period. The relative poverty position of the elderly has not improved in the United Kingdom as it has in the United States.

Canada has had minor but noticeable steepening in its profile across the working ages, with a decline in relative income and a rise in 'young households' poverty rates. There has been no real effect on poverty among households over age 30. The relative income position of older Canadian households has improved dramatically. Canadian old-age benefits have become highly targeted on the low-income elderly, producing a large decline in poverty status, from levels of 24-40 percent in 1975 to 4-6 percent in 1994.

The age/ADPI profile in Sweden has been steepening across the working ages, with the peak income moving to older ages during the 1980's. The targeting of transfers upon low-income working-age families explains why there has been no discernible change in poverty among this group. Unlike the other countries, the profile among older households has actually steepened in Sweden between 1981 and 1992. Although the relative position of older households improved in Sweden between 1975 and 1981, it has not improved since. Indeed, poverty among older households has increased in Sweden and is similar to that in Canada, though still much lower than in the United States or the United Kingdom.

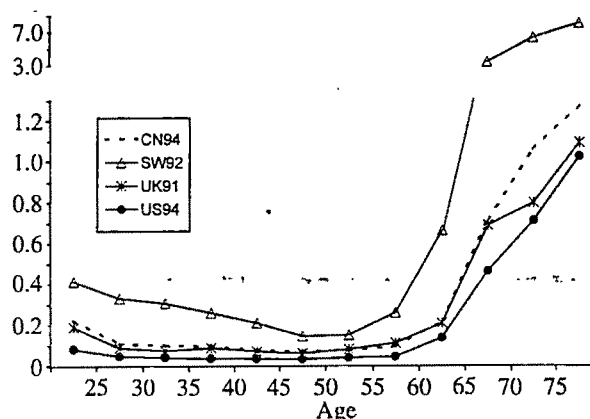


FIGURE 2. RATIO OF SOCIAL-TRANSFER INCOME TO MARKET INCOME OF HOUSEHOLD HEAD BY AGE IN THE EARLY 1990's FOR FOUR NATIONS

Note: The ratio of social-transfer income to market income is computed as the ratio of the sums of total social transfers to market income within each age-of-head group.

We conclude that, based on these analyses, the young appear to be doing worse now relative to 20 years ago in all nations examined here, while the aged are a mixed group. The aged have made continued progress in Canada and the United States and less in Sweden and the United Kingdom.

III. The Role of Social Expenditures via Income Transfers

In this section of the paper we examine the mix of social transfers compared to market income for these populations. Rather than show social transfers as a fraction of total income, we present the ratio of transfers to market income. The reason for this presentation is to show the trade-off between income support from government and income support from the market (earnings, capital income, private pensions, etc.) over time and across nations.

We begin by noting the overall ratios of transfers to market incomes in the most recent year, ranging from 10.6 percent in the United States to 48.8 percent in Sweden, with Canada (16.7 percent) and the United Kingdom (16.3 percent) between these two. Given these differences in support levels, Canada, the United Kingdom, and the United States have the most similar patterns of transfers to market income by age group in the early 1990's (Fig. 2).

Among the elderly, the ratio of transfers to other support (earnings, pensions, capital income) rises rapidly beyond age 60, peaking at roughly a 1:1 ratio by age 75 in the United States, with slightly higher ratios in Canada and the United Kingdom. The United States is the most frugal transferor at all ages; Canada and the United Kingdom are very similar up to older ages, where Canada spends a bit more than does the United Kingdom. Sweden is literally off the chart at beyond age 60, with ratios of 7 or 8 kronor of transfer to each kronor of market income at older ages.

These expenditures are mirrored somewhat in the poverty rates in Figure 1, up until the 55–59 age group. Beyond age 60 social expenditure and antipoverty performance differ greatly. The United States, which spends the least, and Sweden, which spends the most, have predictable outcomes. The most interesting difference is between Canada and the United Kingdom, which spend roughly the same overall amount on social transfers. In Canada, old-age expenditures have been increasingly targeted toward those who would otherwise be poor. A strong income-tested old-age supplement to the Canadian Social Security pension has produced Canadian old-age poverty rates at or below Swedish rates in recent years (Card and Richard Freeman, 1993; Keith Banting, 1997). In contrast, the United Kingdom lower-tier benefit to the social retirement system is ineffective in reducing old-age poverty, while the upper tier of the system produces high benefits for the nonpoor aged. Thus, it is not just what you spend, but how well you spend it that matters in terms of antipoverty performance.

Trends in transfers relative to market incomes differ widely across nations. The relative level of transfer in the United States is remarkably constant over the 1974–1994 period in both level of outlay and pattern. Because poverty rates have risen from 16.4 percent to 19.1 percent over this period, the transfer system has lost a bit of its antipoverty effectiveness, but the overall consistency for each age range is remarkable.

In contrast, Sweden's ratio of transfer to market income has risen from 26.6 percent to 48.8 percent over the period, with most of the increase going to those over age 65. Poverty

has remained at 6.7 percent overall for Sweden, but the cost of benefits for the aged has skyrocketed, owing to the relatively large fraction of Swedes who are now aged 65 or older.

In Canada, social expenditures grew from 1979 to 1994 and were targeted at the old and at the young. In particular, we note a large increase in the ratio of transfer to market income at older ages. This matches other recent evidence that Canada's male labor-force participation rate has declined more rapidly over this period than in any of the other nations studied here (Organization for Economic Co-operation and Development, 1996).

The United Kingdom offers the most interesting patterns of change in transfers across the life cycle over time. After a more than doubling of the ratio of transfers to market income from 9.7 percent to 19.0 percent from 1974 to 1979, the ratio fell back to 16.3 percent by 1991. While the United Kingdom did not receive much of an antipoverty dividend from higher transfers between 1974 and 1979, the lowering of transfers from 1979 to 1994 was associated with a rise in overall poverty from 9.2 percent to 14.6 percent. Most of the decline in transfer beyond 1979 was absorbed by the aged, whose poverty rates rose with the partial privatization of the U.K. social retirement system and the relatively meager level of lower-tier assured old-age benefits that these reforms brought with it.

IV. Discussion and Future Research

Based on four nations, the LIS data offer a rich opportunity to explore the changing fortunes of age groups (and of cohorts) over time. Our next steps are to look more closely beneath these trends to separate the changes in living arrangements, retirement, marriage, and fertility which underlie the patterns we observe. Clearly, most of the change we observe comes from economic and demographic changes at young and old ages, two groups to whom we briefly turn.

Card and Lemieux (1997) find that, at earlier ages, Canadian youth responded to the deteriorating Canadian labor market of the late 1980's and early 1990's by living with their parents, while in the United States fewer youth returned to the parents' nest. These findings

are entirely consistent with the greater steepening of the age-income profile in the United States than in Canada. The earnings-related incentives for added years of education and their effect on emerging patterns of labor-market participation are playing themselves out in each of these nations. The implications for the cost of education, choice of living arrangements, age at first marriage, and fertility need to be further examined within each of these nations in the Easterlin tradition.

A second line of further inquiry deals with the transfer costs of an aging society. The ratio of transfers to market income have risen dramatically in Canada and in Sweden from age 55–59 onward in recent years. Like the United States, Canada and Sweden face a steep future cost of an aging society. Sweden is, in fact, well into the aging of their population, while Canada and the United States are not going to feel the pressure for another decade or longer. However, the trend toward early retirement (and early take-up of social transfers) in Canada is much greater than in the other nations considered here. The United States has, in fact, halted its decline in labor-force participation for older men and is experiencing a slight reversal. But while the Social Security spending implications of more work at older ages seem rosy for the United States, the effectiveness of the old-age transfer system in preventing old-age poverty in the United States is far less than in Canada (or in Sweden). The United Kingdom has largely managed to avoid a future old-age retirement cost crisis with its two-tiered scheme whereby only the first tier depends on tax revenues to fund transfers, and the second tier is largely self-funded by retirees' own contributions. However, a pattern of high and rising old-age poverty rates in the United Kingdom also emerges from this analysis.

As the United States faces a reform of its old-age social transfer system, there is much to be learned from cross-national comparisons of transfer cost, social policy design, and outcomes such as old-age poverty rates. Can the United States reform its system to encourage greater market incomes (via higher savings rates, better occupational pensions, and increased market earnings) at older ages? Can it avoid the high social transfer costs of old age

which will beset many other nations? The British seem to have achieved this avoidance of higher taxpayer cost, albeit at the price of rising old-age poverty rates. And so, unlike Britain, can the United States reform its old-age Social Security system to avoid an increase in old-age poverty, and perhaps like Canada, further reduce economic misery and poverty in old age? Further cross-national analyses along the lines suggested here may reveal a path for the United States Social Security system that produces both a slower growth in overall outlays and greater antipoverty effectiveness.

REFERENCES

- Banting, Keith. "The Social Policy Divide: The Welfare State in Canada and the United States," in Keith Banting, George Huberg, and Richard Simeon, eds., *Degrees of freedom*. Kingston, ON, Canada: McGill-Queens University Press, 1997, pp. 267–309.
- Card, David and Freeman, Richard, eds. *Small differences that matter*. Chicago: University of Chicago Press, 1993.
- Card, David and Lemieux, Thomas. "Adapting to Circumstances: The Evolution of Work, School, and Living Arrangements among North American Youth." National Bureau of Economic Research (Cambridge, MA) Working Paper No. 6142, August 1997.
- Duncan, Greg J.; Boisjoly, Johanne and Smeeding, Timothy M. "Economic Mobility of Young Workers in the 1970s and 1980s." *Demography*, November 1996, 33(4), pp. 457–509.
- Easterlin, Richard A. *Birth and fortune*, 2nd Ed. Chicago: University of Chicago Press, 1987.
- Gottschalk, Peter and Smeeding, Timothy M. "Cross-National Comparisons of Earnings and Income Inequality." *Journal of Economic Literature*, June 1997, 35(2), pp. 633–86.
- Organization for Economic Cooperation and Development. *Aging in OECD countries: A critical policy challenge*, Social Policy Studies No. 20. Paris: Organization for Economic Cooperation and Development, 1996.

Relative Cohort Size and Inequality in the United States

By DIANE J. MACUNOVICH*

Can an increase in relative cohort size (RCS; the ratio of young to older adults in the population) have beneficial economic effects? Economists have long recognized the potential adverse *supply* effects of increased RCS on the relative wages of the young, because of imperfect substitutability between older and younger workers in the labor market.¹ But are there offsetting beneficial *demand* effects on wages?

It is hypothesized that changes in domestic consumption—and in the induced investment generated by that consumption—have resulted from the sharp changes which have occurred in various age groups in the population in the postwar period. The passage of the baby boom, and then the baby bust, into the labor-market and household-formation stages was characterized by a number of “spikes” when growth surged and then fell dramatically. In a market as finely tuned to changes in “underlying fundamentals” as the U.S. market, such marked fluctuations are likely to have caused strong ripple effects through investment and consumption multipliers.

In addition, the baby boomers, as children in their parents’ households, contributed to significant changes in consumption as a proportion of household income, as they grew from toddlers to teenagers: parents spend over 15-percent more, out of a given income, for older children relative to those aged 0–5 (Macunovich, 1998). These changes contributed to the strong growth in the economy in the 1960’s, and then to the dramatic falloff in the 1970’s as the boom in children turned into a bust.

I. Rethinking Relative-Cohort-Size Measures

It is fairly typical in analyses of relative cohort size (RCS) effects, to develop cohort size measures using labor-force data. Many have used ratios of the number of workers in each education–experience cell relative to the total number with that level of education, as in Finis Welch (1979). Murphy and Welch (1992) take this to an extreme by calculating not simply the number of workers, but the number of *hours worked*, by members of each education–experience cell. This type of calculation ignores any potential endogeneity of hours and weeks worked, educational attainment, and even labor-force participation rates, with respect to RCS.

This analysis attempts to measure both aggregate supply and demand aspects of population change using two fairly straightforward population (not labor-force) ratios. The first, the general fertility rate (GFR) in an individual’s year of birth, referred to as birth-cohort size, remains constant for all members of a given birth cohort throughout their lifetimes.² The second, the ratio at time *t* of population in their early twenties relative to prime-age adults, referred to as current cohort size, is a measure of current conditions and is the same for all individuals at a given point in time. Its effect, however, is expected to vary by level of experience.

Members of a large birth cohort are expected to fare worse due to the excess supply of labor they provide as they pass through the labor market, a standard concept. But at any point in time, current cohort size will mediate the effects of individuals’ own birth-cohort size because of positive aggregate demand effects: the overall wage structure will rise and fall in response to the current ratio. This positive aggregate demand effect will be reflected

* Maxwell Center for Policy Research, Syracuse University, Syracuse, NY 13244. I am grateful for the many types of support I have received from the people at the Maxwell Center, financial support through an NIA Fellowship, and Richard Easterlin’s inspiration and support. I also thank Lee Lillard for providing my first opportunity to work with the CPS on “youth labor markets.”

¹ See Macunovich (1998) for detail on the literature.

² The general fertility rate is the number of births per 1,000 women aged 15–44 in a given year.

in the level of the current-cohort-size measure, and also in the rates of change in the two measures, birth- and current-cohort size.

These rates of change are referred to here as "derivatives," although technically they are simply first differences. It is assumed that the adverse effects of large birth-cohort size on individuals born on an upswing in fertility (positive first derivative) will be ameliorated to some extent by the positive effects of increasing cohort size on the demand for goods and services and that those born near peaks in the fertility rate (negative second derivative) will benefit from the sharp changes they induce in the composition of demand.

The U.S. general fertility rate, lagged 20 years, and a measure of current cohort size (population aged 20–22 relative to population aged 45–49) are presented in Figure 1, together with their first derivatives. The pattern in the first two panels is notable for the contrast between strong positive values in the 1950's, 1960's, and early 1970's, and strongly negative values in the following 20 years.

Even more striking is the pattern established in the second derivative of the lagged GFR (bottom panel). This derivative differentiates between peaks and troughs (indicates turning points), whereas the first derivative differentiates between periods of increase and periods of decline. The second derivative is a harbinger of approaching decline in the proportion of young people in the population, and in that sense its historic pattern is striking for its conformance with changes in the economy over the last century.

Since at least 1920 all of the troughs in the second derivative have coincided with recessions some 20 years later. The fact that these troughs were less severe prior to World War II, despite the severity of recessions in that period, suggests that monetary and fiscal policies since World War II now cushion the effects of large population fluctuations.

II. Data and Methodology

The attempt throughout this analysis has been to ensure the comparability of results with other studies of cohort-size effects in the labor market. The data set reproduces (and updates through 1996) that used in Kevin Murphy and Finis Welch (1992), with a rel-

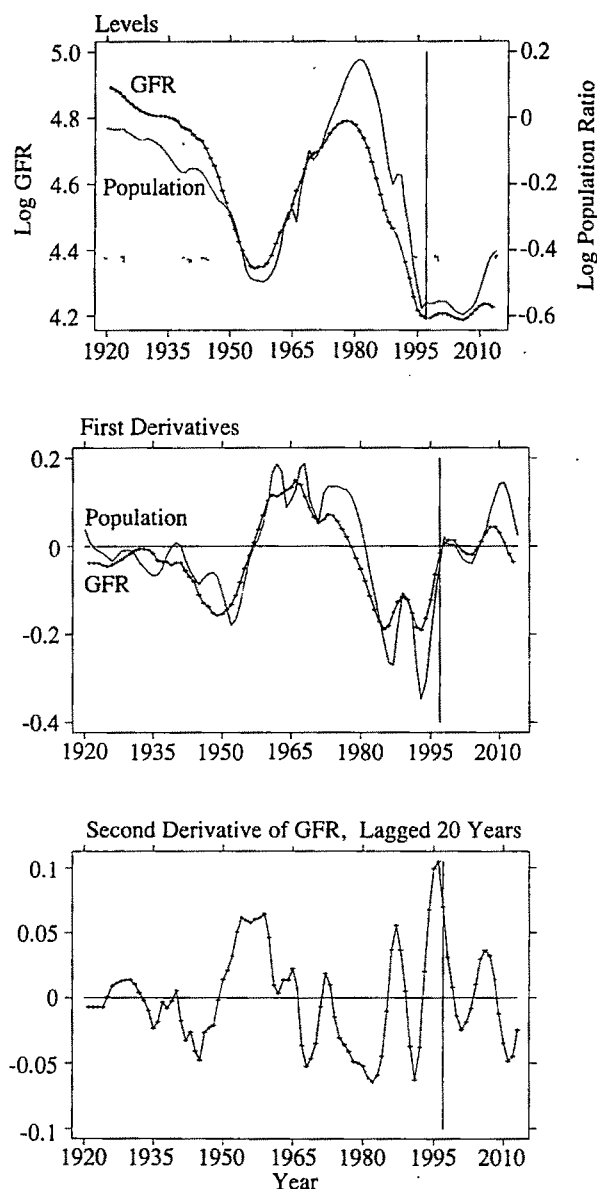


FIGURE 1. GENERAL FERTILITY RATE (GFR), FIVE-YEAR MOVING AVERAGE, LAGGED 20 YEARS, AND CURRENT POPULATION RATIO ($[AGES\ 20-22]/[AGES\ 45-49]$)

ative cohort size (RCS) variable similar to the one used there and, for example, in Welch (1979). That is, the data were taken from the annual CPS for the years 1964–1996,³ and the wage sample, as in Murphy and Welch (1992), was restricted to white civilian men who worked full time at least 40 weeks during

³ The years 1964–1995 were taken from the uniform version of the March Current Population Survey prepared by Unicon; 1996 was taken from the original CPS tapes.

the year, excluding men with self-employment income and men whose wages were imputed.⁴

In order to control for potential forces other than RCS, three macroeconomic variables were included, all in logged form: the annual change in total real GDP; the real per capita level of the current durable-goods trade deficit (imports minus exports); and the ratio of 20–24 year olds relative to the total active military each year. The first of these was included, despite its hypothesized endogeneity, to test for its residual effect in the presence of the RCS variables. For further details on data and methodology, see Macunovich (1998).

III. Results

Table 1 presents partial results (normalized coefficient estimates), from a number of different models to compare the effects of the three RCS measures. Coefficients on the birth and current-cohort-size measures are highly significant and display the expected signs consistently in all model formulations. The magnitude of their total effect is more than ten times that estimated using the Welch RCS measure, which is positive and only weakly significant.

The GDP-change variable produces a counterintuitive negative effect on wages, in the absence of cohort-size variables [model (i)] and when included with the Welch variable [models (ii) and (iii)] and with the birth-cohort-size variable on its own [models (v)–(vii)]. Its effect becomes positive, however, in the presence of the current-cohort-size variable [specifications (ix) and (x)], although in Table 2 it loses its significance, except for those with 13–15 years of education, in a model which allows all coefficients to vary by education level.

Thus, consistent with the hypothesis, the current-cohort-size variable produces a strong

TABLE 1—ESTIMATED COEFFICIENTS USING WELCH, BIRTH, AND CURRENT COHORT-SIZE VARIABLES

Independent variable	Specification			
	(i)	(ii)	(iii)	
Welch cohort size		0.009 (3.2)	0.007 (2.8)	
$\Delta \log(\text{GDP})$	-0.019 (-8.2)	-0.019 (-8.2)		
Independent variable	Specification			
	(iv)	(v)	(vi)	(vii)
Birth cohort size:				
Level	-0.047 (-23.6)	-0.043 (-21.3)	-0.046 (-23.0)	-0.074 (-21.8)
First derivative			0.042 (18.3)	0.047 (20.1)
Second derivative				-0.034 (-10.6)
$\Delta \log(\text{GDP})$		-0.020 (-8.5)	-0.018 (-7.8)	-0.017 (-7.5)
Independent variable	Specification			
	(viii)	(ix)	(x)	
Birth cohort size:				
Level			-0.067 (-15.6)	
First derivative			0.055 (17.1)	
Second derivative			-0.022 (-6.6)	
Current cohort size:				
Level	0.066 (13.8)	0.052 (10.8)	0.040 (7.8)	
First derivative	0.097 (20.1)	0.080 (14.9)	0.018 (3.1)	
$\Delta \log(\text{GDP})$		0.008 (3.3)	0.007 (2.9)	

Notes: The table presents partial results with normalized coefficients; the dependent variable is $\log(\text{hourly wage})$. Numbers in parentheses are t statistics. The birth-cohort-size measure is GFR in each individual's year of birth. The current-cohort-size measure in (viii)–(x) is regional population aged 20–22/45–49. All regressions included a time trend and controls for education (6), experience (17), and state (21). Specifications (viii)–(x) included full sets of interaction terms between experience levels and the two current-cohort-size variables. Specifications (i), (ii), (v)–(vii), (ix), and (x) included controls for size of military and trade deficit. There were approximately 145,000 observations.

⁴ Finis Welch kindly provided the algorithms used in imputing hours and weeks worked for the years prior to 1976, which were used in calculating average hourly wages in the wage sample, and total hours worked in the employment sample. Log hourly real wages were averaged in approximately 145,000 weighted experience-education-state-year cells.

TABLE 2—ESTIMATED COEFFICIENTS BY EDUCATION LEVEL, USING BIRTH AND CURRENT-COHORT-SIZE MEASURES

Independent variable	12 years (basic effect)	Education level		
		<12 years	13–15 years	16+ years
Birth cohort size:				
Level	-0.078 (-11.8)	-0.671 (-4.3)	—	-0.147 (-0.9)
First derivative	0.059 (11.3)	-0.029 (-6.8)	—	-0.016 (-3.4)
Second derivative	-0.020 (-4.1)	-0.015 (-3.5)	—	—
Current cohort size:				
Level	0.093 (12.1)	0.045 (4.5)	-0.028 (-4.2)	-0.069 (-9.4)
First derivative	0.039 (4.3)	0.030 (3.6)	-0.011 (-1.6)	-0.036 (-4.6)
$\Delta \log(\text{GDP})$	0.005 (1.4)	—	-0.010 (-2.4)	—

Notes: The table presents partial results of a pooled regression using model (x) of Table 1, but with full sets of interaction terms for each of three education levels (<12 years, 13–15 years and 16+ years). Table entries are normalized coefficients, with *t* statistics in parentheses; the dependent variable is $\log(\text{hourly wage})$. Regression included a time trend, controls for size of military and trade deficit, and controls for experience (17) and state (21), with interaction terms between experience levels and the two current-cohort-size variables. There were approximately 145,000 observations. The current-cohort-size measure is regional population aged 20–22/45–49. A dash indicates a zero estimate for the differential effect.

positive effect on wages, both in its level and in its rate of change, which is highly correlated with GDP growth. An individual's own birth-cohort-size measure produces a strong negative effect which is ameliorated on the leading edge of any increase (first derivative positive) and near peaks in the GFR (second derivative negative). In addition, in results not presented here because of space constraints, interaction terms between the current-cohort-size measure and experience dummies are also significant and display a differential positive effect which most benefits prime-age workers.

The fit of model (x) to observed data is illustrated in Figure 2, for groups at opposite ends of the experience spectrum: those with 1–10 and 41–45 years of experience. The younger workers' hourly wage is shown in a ratio to that of prime age men with 26–35 years of experience, while the older workers are compared with the group with 36–40 years of experience, in order to get a sense of their

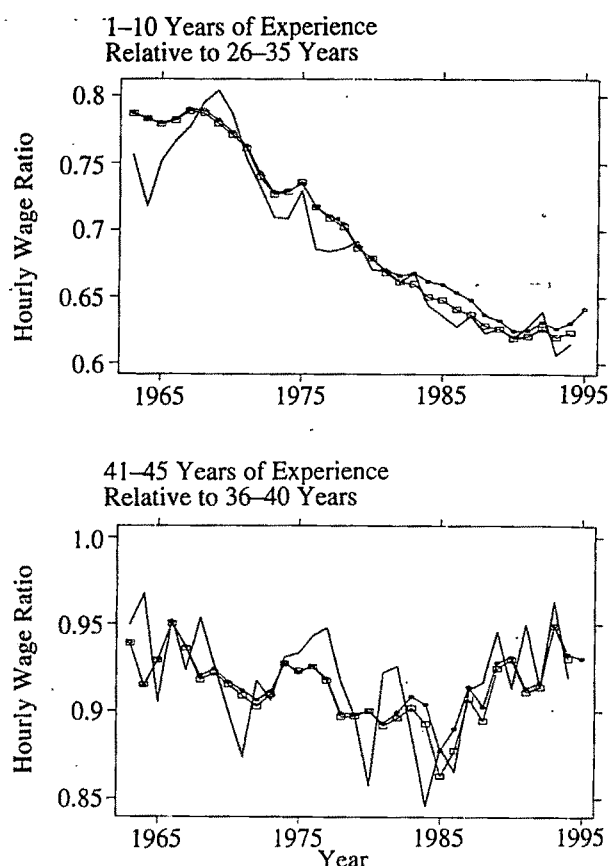


FIGURE 2. OBSERVED (—), SIMULATED (—○—), AND PREDICTED (—□—) HOURLY WAGE RATIOS, USING MODEL (x) FROM TABLE 1

wage progression as they pass through their final years in the labor market.

Each of the graphs in Figure 2 presents the observed time pattern of the wage ratio, together with the respective model's predicted value using all information on the historic pattern of the macroeconomic variables, and a "simulated" value obtained by holding all variables, other than those measuring cohort size, constant at their 1980 levels. The close match between the simulated and observed values in almost all cases indicates that nearly all of the relative wage movement over the past 33 years has been due to changing RCS effects. Even the absolute levels of real wages at each experience level have reflected fairly closely the effects of changing RCS, as illustrated in Figure 3.

However, Table 2 shows that there are marked differences in cohort-size effects by education level, and Figure 4 suggests the

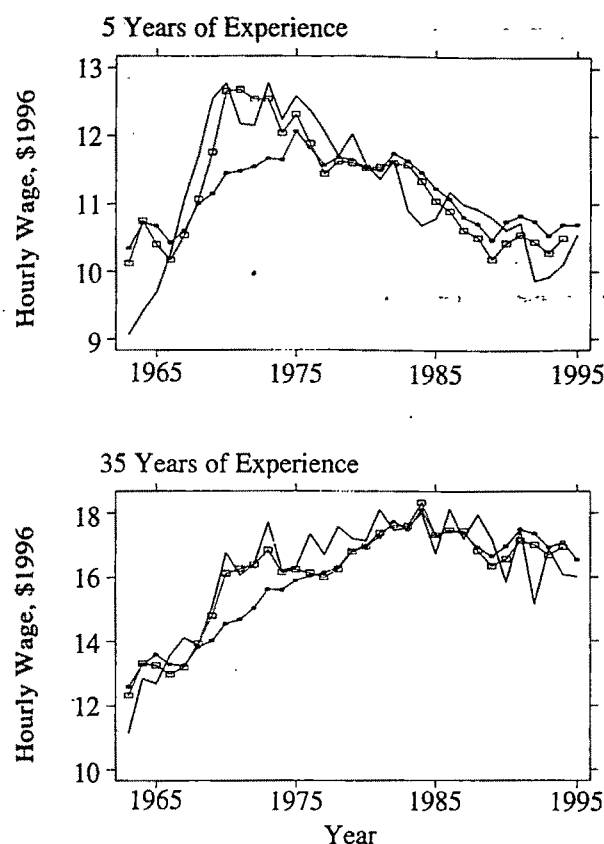


FIGURE 3. OBSERVED (—), SIMULATED (—○—), AND PREDICTED (—□—) HOURLY WAGES (\$1996)

strong differential effects of the macroeconomic control variables included in the regressions (reported in detail in Macunovich [1998]). The differences between the predicted and the “simulated” curves in Figure 4 represent the effects of these other variables, which lowered the college premium in the late 1960’s and raised it in the 1980’s and 1990’s, relative to the path it would have followed based on changing cohort size alone.

IV. Discussion

The work presented here has attempted to test the hypothesis that changing demographic structure has been a major factor in the changes in relative wages that have occurred over the last 30 years, leading to the observed sharp decline in the wages of young adults and those approaching retirement, relative to prime-age workers, as well as to the decline and then steep increase in the wages of the college-educated relative to high-school grad-

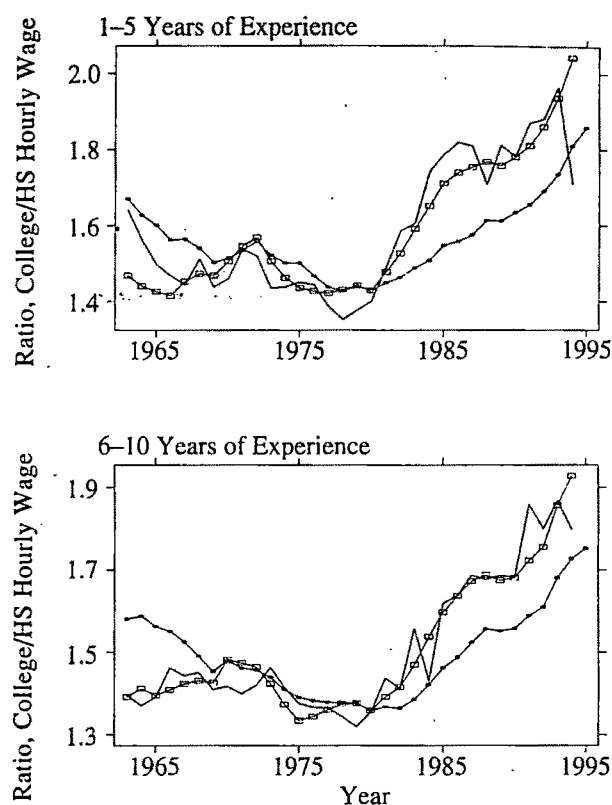


FIGURE 4. OBSERVED (—), SIMULATED (—○—), AND PREDICTED (—□—) COLLEGE WAGE PREMIUM

uates. The belief is that studies that have attempted to quantify such effects in the past have erred both in their method of representing age-structure changes in their models (their choice of relative cohort size measures) and in their failure to allow for the possibility that changing age structure might have strong aggregate demand as well as aggregate labor-supply effects in the economy.

The analysis has identified pronounced effects of changing age structure on wages: almost all of the change in the experience premium over the past 30 years (younger and older relative to prime-age workers) and a significant portion of the change in the college wage premium can be explained solely as a function of changing age structure. Other factors, such as the Vietnam conflict and changing levels of international trade, have indeed played some role, but not to the extent that has been assumed. Figure 5 illustrates two patterns of wage–experience profiles (simulated on the basis of changing demographic structure and observed) at three different points in time: 1965, 1980, and 1995.

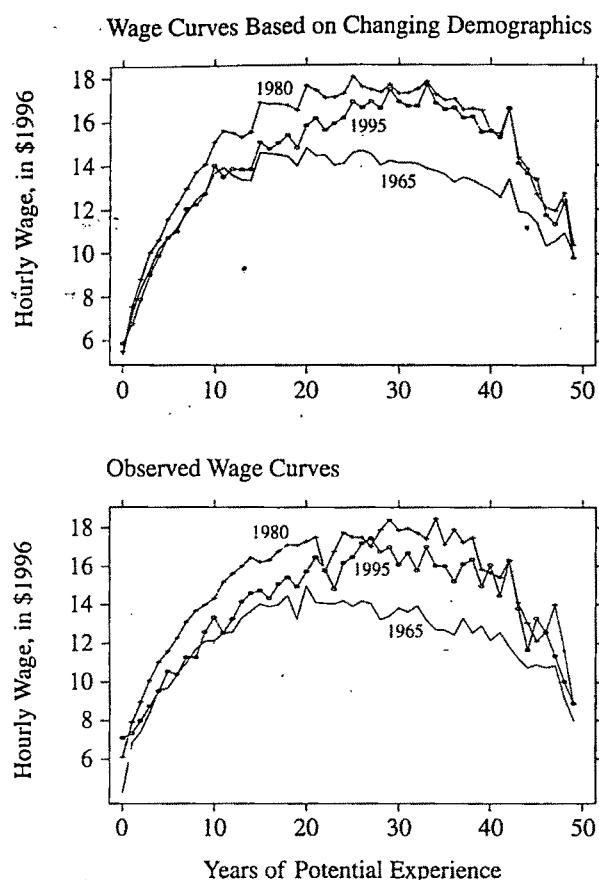


FIGURE 5. WAGE-EXPERIENCE PROFILES: SIMULATED (TOP) AND OBSERVED (BOTTOM)

The similarities are obvious, with the extremely low and relatively flat profile in 1965, the high more sharply peaking curve in 1980, and then the fall to a flatter curve in 1995—but one with a fairly pronounced hump favoring those with 30–40 years of experience, relative to the pattern in 1965.

The implications of these findings are wide-ranging. The baby boom was in fact a mixed curse—or blessing, depending on one's point of view. Those born in the first half of the boom fared poorly, but only in relative terms, because the older members of the population fared so well in an expanding economy fueled by the expenditures of the baby boomers' parents, and later by the boomers themselves. Those who

were born as the flow ebbed have been hit hardest because they were still relatively large cohorts and had to compete with peak boomers who were still trying to find their appropriate niche in the labor market, and they emerged when economic growth was weakened by their own declining numbers.

One should not conclude from these findings that a baby boom is the cure for all of an economy's ills. The benefits of lower fertility have been well documented. But these findings add a new twist to the theory of demographic transition: a potential hurdle further down the line as the full effects of a slowdown in population growth hit the newly expanding economy. And they generate a host of new questions. Is the wave of "currency crises" that has hit several of the growing new economies a result (at least in part) of this type of deflation (i.e., a lagged response to their falling birth rates 20–30 years earlier)? Did the U.S. demographic transition lead to the Depression, which in turn generated the baby boom? Has the marked decline in the retirement age been at least in part a function of the baby boom's passage into the labor market, and the subsequent drop in relative earning potential for older workers? And, if there are indeed larger macroeconomic implications of changing population age structure, what role, if any, should be played by society in evening out the very unevenly distributed spoils?

REFERENCES

- Macunovich, Diane J. "The Fortunes of One's Birth: Relative Cohort Size and the Wage Structure in the U.S." Mimeo, Syracuse University, 1998.
- Murphy, Kevin and Welch, Finis. "The Structure of Wages." *Quarterly Journal of Economics*, February 1992, 107(1), pp. 285–326.
- Welch, Finis. "Effects of Cohort Size on Earnings: The Baby Boom Babies' Financial Bust." *Journal of Political Economy*, October 1979, 87(5), part 2, pp. S65–97.

Intergenerational Transmission of Health

By DENNIS AHLBURG*

Social scientists have long been interested in the intergenerational transmission of income and socioeconomic status because if these attributes are sufficiently intergenerational this would be inconsistent with the ideal of equal opportunity, that is, that a young person's economic future should not be determined by his or her origins. If the intergenerational transmission of income or status is sufficiently large and reflects inequality of opportunity, then government intervention may be called for to increase both equity and economic efficiency.

The statistical evidence for the United States and other developed countries has generally indicated an intergenerational transmission of income of about 0.2. However, more recent studies have found that, after correcting for measurement error, unrepresentative homogeneous samples, and transitory fluctuations in income, the intergenerational transmission is much higher, at least 0.4. (Gary Solon, 1992; Jere Behrman et al., 1995).¹ Thus intergenerational mobility in the United States and possibly also in other developed countries is dramatically lower than is commonly thought and has probably been decreasing since 1980. Intergenerational mobility in developing countries is probably lower than in developed countries, but lack of data limits comparisons. The findings of higher intergenerational correlation than previously believed also resolves a seeming inconsistency between studies that have found childhood background to be an important determinant of adult socioeconomic success and the studies of intergenerational transmission of income (Behrman et al., 1995 p. 241).

These studies have not decomposed the estimated intergenerational correlations into

causal components. Such a decomposition would help to identify factors that promote or retard mobility and also identify possible paths for government intervention. A further decomposition could identify the extent to which the causal components reflect genetic or environmental influences, that is, address the nature versus nurture debate. Human-capital theory suggests that education and health are prime suspects in endowments (Behrman et al., 1994). I will briefly consider evidence on the intergenerational transmission of education before considering evidence on health and longevity and the allocation of resources within the household, in particular whether such allocation is compensating or reinforcing of initial endowments.

I. Estimates of the Intergenerational Transmission of Education

Economists have studied investment in education using the model of Gary Becker in which demand for education varies with an individual's ability or endowments and the supply varies with family wealth and earnings. Behrman et al. (1995) use this model to estimate the intergenerational transmission of education and the contribution of genetics to the acquisition of education. They find intergenerational correlations on the order of 0.3–0.4. Interestingly, they find significant differences in educational-attainment correlations between fraternal twins and other siblings which may reflect different family and societal environments for children born at different times and differences due to birth order and birth spacing. Evidence is presented that these differences reflect parental preferences underlying schooling investments. Thus there are notable differences in the intragenerational transmission of education, and depending upon which child is used in the calculation of the intergenerational coefficient, the intragenerational investment decisions will affect the estimate of the intergenerational transmission.

* Industrial Relations Center, University of Minnesota, Minneapolis, MN 55455.

¹ The correlation may actually be higher since the sample excludes those with zero earnings and thus those in poor health.

Behrman et al. (1995) attribute about 80 percent of the variance in educational attainment to genetic influences. Consequently, they find that the correlation in educational attainment in families falls as genetic distance rises. The estimate of heritability is substantially larger than those found in earlier studies. Although the variability attributed to environmental factors is relatively low, changes in environment can have large effects. For example, a child whose father is a manager or professional rather than an unskilled worker has one more year of education, and each extra sibling a child has reduces his or her educational attainment by 0.15 year. Such differences in human capital can translate into large differences in income. The value of an extra year of education varies over time, being relatively low in the 1970's but rising in the 1980's and 1990's. The value of education today is such that education distinguishes between those who will be economically successful and those who will not.

The research on education points to significant intergenerational transmission which likely contributes to the intergenerational transmission of earnings and income. It indicates some scope for interventions which may increase both efficiency and equity and also points to possibly significant intragenerational differences in education, which may also indicate the desirability of policy interventions.

II. Estimates of the Intergenerational Correlation of Health

While all seek health, few agree on how to measure it, perhaps because it is really multidimensional with different aspects of health having different effects on well-being, productivity, and other labor-market outcomes. In addition, health changes over time. In the economics literature, perhaps the most common measures of health are anthropometric measures: height and weight. These have been particularly important in development and economic history. For example, Robert Fogel (1992) showed height to be related to adult mortality and morbidity in historical series. In addition to height and weight and another objective measure, nutrient intake, economists have also used subjective measures: respon-

dent's self-reports of their health status. Indeed, in the empirical literature, self-reported health status (or parent's report of illness in a child) is arguably the most widely used measure of health, despite being subject to systematic measurement error (John Strauss and Duncan Thomas, 1996). Part of height and weight are genetic, and even self-reported measures may be partly genetic; however, economists have done little with the genetic component, usually just controlling for parents' height or weight in equations for child's height or weight.

Studies of the intergenerational transmission of health have generally been carried out by genetic epidemiologists who have sought associations between genes and specific diseases such as cancer and Alzheimer's disease. Another line of study has looked at lifespan, which can be considered as the ultimate output of the health production function.² Estimates of the correlation of the lifespan between generations have been used to support a strong genetic component of longevity (Anatoli Yashin and Ivan Iachine, 1997). Estimates of the correlation between lifespans of parents and children have generally been about 0.15–0.3 (Yashin and Iachine, 1997 p. 32). Some studies interpret this correlation as support for a strong genetic component of longevity, while others have argued that it suggests an environmental nature to the lifespan correlation. However, the parent–child correlation provides only an indirect estimate of the genetic effects. The heritability coefficient rather than the intergenerational coefficient measures the contribution of genetic factors to trait variability. In the absence of shared environmental effects, the expected parent–child correlation is half the narrow-sense heritability (Matthew McGue et al., 1993 pp. B237–38). Geneticists distinguish between two types of environmental effects: shared environmental effects that are due to common rearing factors such as parent's socioeconomic status, childhood nutrition, and the like; and nonshared environ-

² The objective and subjective health measures are not unrelated. Self-perceptions of general health status are believed to be good predictors of mortality (see Strauss and Thomas, 1996 p. 16).

mental effects that include all environmental factors not shared by individuals who were reared together, such as accidents, and adult nutrition. However, if parents discriminate among their children in the allocation of food and health care, factors that are generally taken to be part of the shared environment, the estimates of the impact of environmental factors may be biased. If such discrimination among children exists, as evidence suggests, then the size of the intergenerational correlation in lifespans may depend upon which child is used in the calculation. Indeed, Morton Glasser (1981) found that the father-son correlation was the highest followed by the son-mother correlation.

In an attempt to obtain more reliable estimates of the genetic and environmental components of longevity, researchers turned to data from twins. Estimates of the correlation in lifespans for identical (MZ) and fraternal (DZ) twins are generally 0.3 and 0.2, respectively (Yashin and Iachine, 1997 p. 32). Using uncensored Danish twin lifespan data, McGue et al. (1993) concluded that lifespan is moderately heritable: they calculated an estimate of broad-sense heritability of 0.23 (broad-sense heritability is the proportion of trait variance that is associated with all genetic influences, both nonadditive effects, which do not contribute to parent-offspring resemblance, and additive genetic influences, which do).

In the classic Becker model of human capital, heritability indicates how much inequality there would be in a competitive economy with perfect capital markets and no random events. Economists have argued about the definition and usefulness of heritability, but Behrman et al. (1995 p. 246) argue that a decomposition is useful, not for what it reveals about heritability, but for what it indicates about the variation in common environment, which represents variation in opportunities (although by this it seems that they are referring to what geneticists call nonshared environment). They concluded that about 20 percent of earnings inequality for white males in the United States (and probably more for other groups) is due to inequality of opportunity, and so there are important efficiency and equity gains that can be made by policies designed to eliminate inequality of opportunity.

In their study of longevity of Danish twins, McGue et al. (1993) found little evidence for shared environmental influences on longevity. Genetic factors account for some of the observed twin similarity in age at death, but a major portion is explained by nonshared environmental influences. This finding may be taken to suggest little impact of common rearing on human lifespan. Such an interpretation assumes that shared rearing means equality of treatment. We argue that this is not likely to be the case, particularly in developing countries.

There are a number of features of these estimates that lead us to question their precision: the estimates of the intergenerational correlations treat censoring in different ways (usually ignoring censored observations), and the decompositions into genetic and environmental components require the use of special identifying assumptions. Yashin and Iachine (1997) have suggested that rather than use data on lifespans it may be more beneficial to base models directly on the genetics of susceptibility to disease. This focus is certainly closer to the concept of health that economists concerned with investments in human capital worry about. Yashin and Iachine present a new model that draws from demography, survival models, quantitative genetics, and genetic epidemiology: the correlated-individual-frailty model. The core of the model is the combination of a genetic model of heterogeneity (frailty) with a proportional-hazards survival model. They find that about half of the variability of frailty (susceptibility to disease and death) is due to genetics, about twice the heritability of longevity estimated by McGue et al. (1993). Further, the only significant environmental factors affecting frailty are nonshared effects, as was found by McGue et al. for longevity. That is, differentials across households in nonshared environment (opportunities), for example in income or wealth, contribute substantially to differences in health and longevity. This implies that improvements in efficiency and equity could be achieved by interventions that address such differential environments or opportunities. Further improvements may also be achieved if there are differences in the environments within the household that differentiate among children

on the basis of their age, sex, or some other characteristic.

III. The Further Economic Importance of Health and Longevity

The concept of the planning horizon is central to optimizing behavior in economics. The expected length of the planning horizon affects savings, consumption, insurance, investments in human capital, labor-supply decisions (including retirement), bequests, and a host of other behaviors and decisions. Despite this, there has been relatively little work on how individuals form expectations about the length of the planning horizon, that is, about how long they will live. Given the evidence on the intergenerational transmission of longevity, it would seem that individuals may form their expectations based at least in part on the longevity of their parents and grandparents and perhaps other kin. Is this in fact the case?

Daniel Hamermesh (1985) used survey data to examine individuals' responses on subjective life expectancy and survival probabilities. He found that subjective estimates of life expectancy exceed actuarial estimates by a little less than two years and that the subjective distributions are flatter than the actuarial distribution. Individuals also extrapolate past changes in life tables when forming their subjective horizons, although they tend to be more optimistic than is warranted by the data. In addition, individuals with grandparents or parents who lived to at least 80 years of age had higher subjective life expectancies, while those whose parents or grandparents died before age 60 had reduced subjective estimates of longevity. The impacts of longevity of parents and grandparents were particularly large: one parent surviving to be 80 increase the life expectancy estimate by 1.3–3.0 years, and both parents surviving increased it by 4.2–5.9 years. The impacts of grandparents' and parents' longevity are 2–3 times the impacts based on epidemiological evidence. Information on forebears' longevity also affects the variance of expected longevity such that having early decedents among one's forebears shifts the subjective distribution to the left and narrows it, while having older decedents shifts

it to the right and broadens it. The conclusions from these findings are clear: "empirical studies of life-cycle saving, investment in human capital, and labor supply ignore changing life expectancy and its effects on subjective horizons and survival probabilities at the expense of realism, and with the possible price of incorrect behavioral implications." (Hamermesh, 1985 p. 406). Specifically, individuals whose parents were in poor health or who died early are likely to form an overly pessimistic estimate of their own morbidity and mortality and to underinvest in human capital, while those with healthy and long-lived forebears may do the opposite. Such behavior would reinforce intergenerational transmission of earnings and income.

IV. Intragenerational Transmission of Health and Education

There is substantial evidence of unequal health and education-input allocations within households. This may be on the basis of age, sex, or some other characteristic. Such differential allocations may reflect parental preferences, differential health endowments, the shape of the health production function, or differential returns in the labor market (Strauss and Thomas, 1996). Under some assumptions, differential allocation can reinforce endowment differences, while under others it can compensate for them. Obviously, the outcome of the family's decisions has important implications for the individual distributions of earnings and income, among other variables.

In both developed and developing countries, there is evidence that some children receive favorable treatment. In some countries with strong son preference, evidence exists that girls suffer from direct infanticide, unfavorable allocation of scarce resources, and inferior health care. In other settings, males are sometimes favored or there is no clear preference or a slight preference in favor of females (Behrman et al., 1995). Part of the explanation for different impacts in different settings is Duncan Thomas's (1994) finding that mothers tend to favor daughters in allocating resources, while fathers tend to favor sons, partly reflecting differences in child-

rearing technologies and partly their preferences. The net effect will depend upon the relative importance of the father and mother in resource-allocation decisions. My recent work with Eric Jensen has found that parents in at least some settings favor wanted children over unwanted children in education and health decisions (Jenson and Ahlburg, 1997). In many of these situations, allocation of resources among children is not necessarily made on efficiency or productivity grounds (Behrman et al., 1995) as predicted by the original model of Becker and Nigel Tomes.³

But does differential treatment by parents mimic or oppose nature? Mark Pitt et al. (1990) found that Bangladeshi household health investments favor individuals with greater health endowments, and in studies of U.S. twins, Behrman et al. (1994) found that investments in education tended to reinforce individual-specific endowment differentials. Thus, households tend to follow nature in their investment strategy. It is possible that public investments follow a similar strategy, further reinforcing the inequalities that household decisions produce.

V. Conclusion

Inequalities in earnings and income tend to be correlated across generations at least in part because of intergenerational correlations in education and health. These correlations appear to be larger than is commonly recognized, implying lower intergenerational mobility than is generally assumed. Much of the intergenerational correlations in human capital seem to be associated with genetic and non-shared environmental factors. Within-family investments appear to reinforce endowments rather than compensate for them. Individuals appear to be aware of the intergenerational correlation of health, indeed, to an exaggerated extent. These results suggest a role for policy interventions that may be justified on

equity grounds and also perhaps on efficiency grounds.

REFERENCES

- Behrman, Jere; Pollak, Robert and Taubman, Paul. *From parent to child*. Chicago: University of Chicago Press, 1995.
- Behrman, Jere; Rosenzweig, Mark and Taubman, Paul. "Endowments and the Allocation of Schooling in the Family and in the Marriage Market: The Twins Experiment." *Journal of Political Economy*, December 1994, 6(102), pp. 1131-74.
- Fogel, Robert. "Second Thoughts on the European Escape from Hunger: Famines, Chronic Malnutrition and Mortality Rates," in S. R. Osmani, ed., *Nutrition and poverty*. Oxford: Clarendon, 1992, pp. 243-86.
- Glasser, Morton. "Is Longevity Inherited?" *Journal of Chronic Diseases*, 1981, 34(9-10), pp. 439-44.
- Hamermesh, Daniel. "Expectations, Life Expectancy, and Economic Behavior." *Quarterly Journal of Economics*, May 1985, 100(2), pp. 389-408.
- Jensen, Eric and Ahlburg, Dennis A. "Within Family Resource Pressures and Child Health in Indonesia, Korea, and the Philippines." Program on Population, East-West Center (Honolulu, HI) Working Paper No. 88-15, 1997.
- McGue, Matthew; Vaupel, James W.; Holm, Niels and Harvald, Bent. "Longevity Is Moderately Heritable in a Sample of Danish Twins Born 1870-1880." *Journal of Gerontology*, November 1993, 48(6), pp. B237-44.
- Pitt, Mark; Rosenzweig, Mark R. and Hassan, Nazmul. "Productivity, Health, and Inequality in the Intrahousehold Distribution of Food in Low-Income Countries." *American Economic Review*, December 1990, 80(5), pp. 1139-56.
- Solon, Gary. "Intergenerational Income Mobility in the United States." *American Economic Review*, June 1992, 82(3), pp. 393-408.
- Strauss, John and Thomas, Duncan. "Health, Nutrition and Economic Development." Mimeo, Michigan State University, 1996.

³ Behrman et al. (1995) developed a "separable earnings-transfer" model that allows parents to care differentially for their children and does not predict that parents necessarily invest efficiently in their children's human capital.

Thomas, Duncan. "Like Father, Like Son: Like Mother, Like Daughter. Parental Resources and Child Height." *Journal of Human Resources*, Fall 1994, 29(4), pp. 950-88.

Yashin, Anatoli and Iachine, Ivan. "How Frailty Models Can be Used for Evaluating Longevity Limits: Taking Advantage of an Interdisciplinary Approach." *Demography*, February 1997, 34(1), pp. 31-48.

ON THE ECONOMICS OF GIVING[†]

Transfers, Empathy Formation, and Reverse Transfers

By ODED STARK AND ITA FALK*

The literature on private transfers tends to differentiate between two main transfer motives: exchange and altruism (for a recent review see John Laitner [1997]; for a recent empirical analysis see Donald Cox and Mark R. Rank [1992]). An exchange-driven transfer is *positively* correlated with the income of the recipient; a recipient is better equipped to provide a service (e.g., insurance or support) to a donor when the recipient's income is higher. A higher anticipated return then prompts a higher transfer. This reasoning implicitly assumes the recipient's willingness to provide a service. An altruism-driven transfer is *negatively* correlated with the income of the recipient. The donor cares about the recipient's well-being. A decline in this well-being prompts an infusion of support aimed at raising the recipient's income and consumption. This reasoning explicitly assumes that the donor's attitude toward the recipient is parameterized by an altruism coefficient attached to the recipient's utility in the donor's utility function and implicitly assumes that the recipient's attitude toward the donor is given; indeed, that in the donor's mind or heart it plays no role whatsoever.

In contrast, this paper draws attention to the possibility that altruism and exchange may be

intertwined, and that in a setup ordinarily viewed as altruistically motivated the attitude of the recipient is endogenous. This analytical track is introduced through the inclusion of a recipient's empathy function in which empathy is induced by gratitude. We formalize the donor's decision-making as an optimization problem that incorporates anticipation of the recipient's gratitude. This gratitude is a function of the size of the donation, the recipient's pre-transfer income, and the donor's pre-transfer income. We assume that gratification is expressed through a probable transfer that is valued by the donor. Consequently, lower recipient's income may be *positively* correlated with a *seemingly altruistic* transfer because such an income is associated with a stronger sense of gratitude. Since under well-specified conditions the donor's utility arising from a gratitude-eliciting transfer in our model and the donor's utility arising from a transfer in the standard pure-altruism model correlate *negatively* with the recipient's pre-transfer income, the ability to infer motive from conduct is jeopardized; the two motives give rise to types of behavior that can be observationally equivalent.

Typically, the literature on altruism studies the implications of altruistic links for allocative behavior, consumption transfers, and well-being, taking altruism as a given (see Stark, 1995 Ch. 1), and only rarely does it venture to explain altruism (see Stark, 1995 Ch. 6). Indeed, the questions of how altruism is instilled and what explains its evolution lie at the very frontier of research on preference formation and transfer behavior. We suggest that transfers, along with the conditions under which they are made, affect preferences and hence that altruism can arise as a response to actions rather than be orthogonal to them.

There is an intense interest in gift-making in social anthropology dating back at least to

[†] *Discussants:* Kenneth J. Arrow, Stanford University; Zvi Griliches, Harvard University; Gary S. Becker, University of Chicago.

* Stark: Department of Economics, University of Oslo, P.O. Box 1095 Blindern, N-0317 Oslo, Norway, and University of Vienna; Falk: Kennedy School of Government, Harvard University, Cambridge, MA 02138. Steinar Holden, David M. Kreps, Andreu Mas-Colell, and Atle Seierstad transferred to us very helpful comments and reflections. They earned our empathy. Partial financial support from the National Institute on Aging (grant RO1-AG13037) is gratefully acknowledged.

Marcel Mauss (1966). The literature arising from that interest has apparently turned a blind eye to the argument that return is prompted by gratitude, concentrating instead on the moral obligation of the recipient of a gift to reciprocate and on the social mechanisms that support, indeed mandate, reciprocity.

George A. Akerlof has pointed out that workers may give a gift to their firm by providing work in excess of the minimum work required because they "tend to develop a sentiment for their co-workers" (Akerlof, 1982 p. 550). Harder work could prompt the firm to relax the pressure on workers who are unable to meet the minimum work required. By working "at a speed in excess of work rules, ... if [a worker] has sympathy for other members of the work group, he derives utility from the firm's generous treatment of other members of the group for whom the work rules are a binding constraint" (Akerlof, 1982 p. 552). What then underlies the ensuing "gift exchange" formed between the firm and its workers is co-worker empathy. Why exactly the empathy ("sentiment") arises is not explained; it is *assumed* to evolve.

Jack Hirshleifer (1987) considers how gratitude (an "emotion") guides the response of agent "Second" to the productive allocation of agent "First." A more cooperative productive decision by First raises Second's income. Consequently, Second's ability to react in a grateful way increases, as does his inclination to react gratefully. Being aware of Second's contingent behavior, First alters his allocation away from the "short-sightedly selfish optimum." First is prompted to choose an allocation that is more favorable to Second because Second's gratitude-motivated transfer to First is rising in Second's income. That Second's gratitude can correlate *negatively* with his initial income, indeed, emanate from a low initial income, is not being considered, however.

I. The Model

We model a transfer made through a gift (donation), which is motivated by self-interest rather than by altruism. A gift is a noncontracted good. A disposition to reciprocate can therefore be expected to arise. The intensity of gratitude and the extent of reciprocity are

likely to be affected by the amount of the help, by the recipient's need for help, and by the donor's perceived generosity in providing the help.

A. Utility, Empathy, and Income

Consider two agents, indexed by i or j , and two periods, $t = 0, 1$; superscripts will henceforth denote the agent, and subscripts will denote the period. We can think of two farmers in a village in a less-developed economy, each facing idiosyncratic income as a result of probable ill health or localized crop damage, although the principles discussed here can be generalized to other settings. There is one commodity in the economy denoted by c . The price of c is equal to 1. Each agent has an expected utility function U^i :

$$(1) \quad U^i = E(u^i) \quad i = 1, 2$$

where u^i denotes periodic utility at $t = 1$ and $E(\cdot)$ denotes expected value at $t = 0$. The periodic utility is of the form

$$(2) \quad u^i = (1 - \alpha^i) \ln(c^i) + \alpha^i \ln(c^j) \\ i, j = 1, 2 \quad i \neq j$$

where (c^i, c^j) is the periodic consumption vector and where α^i is the periodic empathy coefficient, $0 \leq \alpha^i < 1$. This coefficient indicates the value agent i assigns to agent j 's well-being, $\ln(c^j)$, in forming his own utility. Empathy encompasses the gratitude of agent i toward agent j . The gratitude, in period 1, depends upon the help in the form of a gift (donation) that an agent had received from his counterpart in period 0; upon the recipient's need for help at the time; and upon the donor's perceived generosity. Let d^j denote the donation from agent j to agent i . Let Y_0^i denote agent i 's pre-transfer income or endowment, $i = 1, 2$. The donation in period 0 cannot exceed the donor's endowment, that is, $d_0^j \leq Y_0^j$. A measure of the recipient's need is $(Y_0^i)^{-1}$, which is negatively related to the recipient's pre-transfer income or endowment. A measure of the donor's generosity is $(Y_0^j)^{-1}$, which is nega-

tively related to the donor's pre-transfer income or endowment. We thus define the empathy function in period 1 as follows:

$$(3) \quad \alpha^i(d_0^j, Y_0^j, Y_0^i) = \frac{(d_0^j)^{g^i}}{Y_0^i Y_0^j}$$

$$i, j = 1, 2 \quad i \neq j$$

where g^i is the elasticity of gratitude with respect to donation. The starting endowments, $Y_0^1 > 1$ and $Y_0^2 > 1$, are given. The idea that, all else equal, a larger donation prompts a stronger gratitude implies that $g^i > 0$. In addition, from the requirement that $\alpha^i < 1$, it follows that $g^i < 1 + \ln Y_0^i / \ln Y_0^j$.¹

Agent i 's period-1 earnings, W_1^i , depend upon the agent's investment in period 0, I_0^i , and upon the uniform-across-agents rate of return $(k - 1) > 0$. There is a distinct possibility that, due to a disaster, the returns to an agent will collapse to the "bankruptcy" level b^i , $i = 1, 2$. The probability that an agent is affected by a disaster is $p > 0$. Thus,

$$(4) \quad W_1^i = (1 - p)kI_0^i + pb^i \quad i = 1, 2.$$

A gift (donation) does not "officially" bind the recipient in any way, but it instills gratitude and, in turn, elicits empathy. The recipient thus feels obliged to help the donor should the donor suffer from a disaster in the subsequent period. (This distinguishes a gift from a loan where the circumstances of the lender do not influence the obligation to repay). However, in the event that the recipient is struck by a disaster, he would not be able to help the donor.

B. The Recipient's Decision

Without loss of generality we take agent 1 to be the agent whose initial endowment is

larger, $Y_0^1 > Y_0^2$. We pursue the case wherein agent 1 is the donor and agent 2 is the recipient. The donation is made in period 0, reciprocity occurs in period 1. Henceforth we suppress both the subscript of d_0^j (we confine our analysis to a donation made in period 0) and the superscript of d_0^j (we consider a donation only from agent 1 to agent 2). Denote by r^p the help the recipient will offer the donor in the event that the donor suffers bankruptcy, $r^p \geq 0$. The recipient does not offer help in the event he suffers bankruptcy himself. The decision variable of the recipient is his offered help, r^p . Since he is given a gift (donation) and is being helped in period 0, gratitude is forged, and thereupon empathy toward the helping donor is sensed. Hence, in period 1, the recipient's utility weighs the well-being of both agents.

The budget constraints of the recipient are given by the deterministic and the stochastic terms below, for period 0 and period 1, respectively:

$$(5) \quad I_0^2 = Y_0^2 + d$$

$$(6) \quad c_1^2 = \begin{cases} kI_0^2 - r^p & \text{with a probability of} \\ & (1 - p)p \\ kI_0^2 & \text{with a probability of} \\ & (1 - p)(1 - p) \\ b^2 & \text{with a probability of} \\ & p. \end{cases}$$

The budget constraints of the donor are given by the deterministic and the stochastic terms below, for period 0 and period 1, respectively:

$$(7) \quad I_0^1 = Y_0^1 - d$$

$$(8) \quad c_1^1 = \begin{cases} b^1 + r^p & \text{with a probability of} \\ & p(1 - p) \\ b^1 & \text{with a probability of} \\ & p^2 \\ kI_0^1 & \text{with a probability of} \\ & (1 - p). \end{cases}$$

¹ We had originally defined the empathy function as

$$\alpha^i(d^j, Y_0^j, Y_0^i) = \frac{k(d^j)^{g^i}}{Y_0^i Y_0^j}$$

where the coefficient k depends on physical units. To simplify the analysis, we have assumed that such units are used that k becomes equal to 1. (It is easily seen that such a choice is possible as long as $g \neq 2$.)

The recipient maximizes his expected utility, written below, taking d as exogenous:

$$\begin{aligned}
 (9) \quad U^2 = & (1-p)^2 \{ (1-\alpha^2) \ln[k(Y_0^2 + d)] \\
 & + \alpha^2 \ln[k(Y_0^1 - d)] \} \\
 & + (1-p)p \{ (1-\alpha^2) \\
 & \times \ln[k(Y_0^2 + d) - r^p] \\
 & + \alpha^2 \ln(b^1 + r^p) \} \\
 & + p(1-p) \{ (1-\alpha^2) \ln(b^2) \\
 & + \alpha^2 \ln[k(Y_0^1 - d)] \} \\
 & + p^2 \{ (1-\alpha^2) \ln(b^2) \\
 & + \alpha^2 \ln(b^1) \}.
 \end{aligned}$$

The first-order condition is:²

$$(10) \quad r^p = \max \{ 0, [\alpha^2(kY_0^2 + kd + b^1) - b^1] \}.$$

The derivatives of r^p (if strictly positive) with respect to the size of the donation, the donor's pre-transfer income, and the recipient's pre-transfer income, respectively, are

$$(11) \quad \partial r^p / \partial d = (\partial \alpha^2 / \partial d)(kY_0^2 + kd + b^1) + \alpha^2 k > 0$$

$$(12) \quad \partial r^p / \partial Y_0^1 = (\partial \alpha^2 / \partial Y_0^1) \times (kY_0^2 + kd + b^1) < 0$$

² The second-order condition is

$$\frac{\partial^2 U^2}{\partial (r^p)^2} = -p(1-p) \left[\frac{\alpha^2}{(b + r^p)^2} + \frac{1 - \alpha^2}{(kY_0^2 + kd - r^p)^2} \right] < 0.$$

$$\begin{aligned}
 (13) \quad \partial r^p / \partial Y_0^2 & = (\partial \alpha^2 / \partial Y_0^2)(kY_0^2 + kd + b^1) \\
 & + \alpha^2 k = -\alpha^2(kd + b^1) / Y_0^2 < 0.
 \end{aligned}$$

Otherwise, if r^p is equal to zero, the derivatives of r^p with respect to d , Y_0^1 , and Y_0^2 are also equal to zero.

C. The Donor's Decision

The donor's utility function depends solely upon his own consumption:

$$\begin{aligned}
 (14) \quad U^1 = & (1-p) \{ \ln[k(Y_0^1 - d)] \} \\
 & + p(1-p) [\ln(b^1 + r^p)] \\
 & + p^2 \ln(b^1).
 \end{aligned}$$

The donor's decision variable is his donation, d . Suppose that the donor knows or correctly estimates the return function (10) of the recipient. Thus we assume that the donor has information about (familiarity with) the recipient that allows him to form rational expectations with respect to the recipient's reaction to a donation. Since gratitude is sensed only by the recipient, we suppress the superscript of g^i ($i = 2$), using henceforth g . The donor's first-order condition is:³

$$(15) \quad d = Y_0^1 - \frac{b^1 + r^p}{p(\partial r^p / \partial d)}.$$

Inserting the explicit values of r^p and $\partial r^p / \partial d$ [equations (10) and (11), respectively], we

³ The second-order condition is

$$\begin{aligned}
 \frac{\partial^2 U^1}{\partial d^2} = & -\frac{1-p}{(Y_0^1 - d)^2} \\
 & - p(1-p)g[k^2(Y_0^2)^2 + 2k^2Y_0^2d + 2b^1kY_0^2 \\
 & + k^2d^2 + 2b^1kd + (b^1)^2 + k^2d^2/g] \\
 & \times [d^2(kY_0^2 + kd + b^1)^2]^{-1} \\
 & < 0.
 \end{aligned}$$

get a quadratic equation in d , the positive root of which is

$$(16) \quad d = \{ pgkY_0^1 + pkY_0^1 - pgkY_0^2 - pgb^1 - kY_0^2 - b^1 + [(pgkY_0^1 + pkY_0^1 - pgkY_0^2 - pgb^1 - kY_0^2 - b^1)^2 + 4pgkY_0^1(pg + p + 1) \times (kY_0^2 + b^1)]^{1/2} \} \times [2k(pg + p + 1)]^{-1}$$

The donor computes d using equation (16). He then computes the respective $r^p(d)$ according to equation (10). If $r^p(d)$ is strictly positive and the second-order condition holds, then d in (16) is optimal. If $r^p(d)$ is zero, the optimal d is given by zero.

From (16) it can be discerned that the donor's optimal donation is positively related to his endowment: all else equal, a higher Y_0^1 elicits a smaller recipient's appreciation of a given donation. To preserve the desired recipient's response, the donor must raise d . Thus, a wealthier donor gives more not because he is more generous per se (not because it is less costly for him to give), but because in the mind (or heart) of the recipient higher wealth depreciates the value of a given donation.

An important feature of a donation (as distinct from a loan or, for that matter, an insurance arrangement) is revealed by the sign of the correlation between the size of the optimal donation and the recipient's pre-transfer income. From (15),

$$(17) \quad \frac{\partial d}{\partial Y_0^1} = \{ [\partial(\partial r^p / \partial d) / \partial Y_0^1] (b^1 + r^p) - (\partial r^p / \partial d)(\partial r^p / \partial Y_0^1) \} \times [p(\partial r^p / \partial d)^2]^{-1} = - \frac{(\alpha^2)^2 k^2}{p(\partial r^p / \partial d)^2} < 0$$

where the second equality draws on equations (11) and (13), and on the relationship $\partial \alpha^2 / \partial d = \alpha^2 g / d$ derived from equation (3). The optimal donation is *negatively* correlated with the recipient's pre-transfer income. Hence, our model and the standard altruism-motivated transfers model give rise to behavioral patterns that are observationally indistinguishable.

II. Conclusion

Our objective has been to illustrate the possibility that transfers rise as the recipient's income declines, without recourse to altruism as the underlying transfer motive. An argument may be made that, since gift-giving in our model is motivated by an exchange consideration, our model is merely the standard exchange model in disguise. As pointed out in the Introduction, the standard exchange model predicts that transfers are positively correlated with the recipient's pre-transfer income. Since our model predicts a negative correlation, our model is anything but a variant of the standard model.

Illumination of a possibility is distinct from a claim to a universal domain. Suppose a colleague who is richer than you walks into your office offering to give you a sum of money. You may consider the behavior odd, be suspicious of the colleague's motives, and decline the offer. Suppose, alternatively, that the colleague walks into your office offering to give you money after you have lost your money. You may consider the behavior noble, accept the money, and feel grateful. By referring to a low pre-transfer recipient's income, not to a *lowered* pre-transfer recipient's income, our model is not capable of distinguishing between these two scenarios. However, the model can be reformulated to facilitate a distinction. We may measure the recipient's need not by $(Y_0^i)^{-1}$ but by $(Y_0^i / \bar{Y}_0^i)^{-1}$ for $Y_0^i < \bar{Y}_0^i$, $(Y_0^i / \bar{Y}_0^i)^{-1} |_{Y_0^i = \bar{Y}_0^i} = 1$ where Y_0^i is the recipient's actual realized income.

The model may plausibly give rise to several implications. Perhaps the optimal distribution (the optimal targeting) of charitable giving or, for that matter, of foreign aid across poor and less-poor beneficiaries could be derived through lines of reasoning akin to the one developed in this paper.

REFERENCES

- Akerlof, George A. "Labor Contracts as Partial Gift Exchange." *Quarterly Journal of Economics*, November 1982, 97(4), pp. 543-69.
- Cox, Donald and Rank, Mark R. "Inter-Vivos Transfers and Intergenerational Exchange." *Review of Economics and Statistics*, May 1992, 74(2), pp. 305-14.
- Hirshleifer, Jack. "On the Emotions as Guarantors of Threats and Promises," in John Dupré, ed., *The latest on the best: Essays in evolution and optimality*. Cambridge, MA: MIT Press, 1987, pp. 307-26.
- Laitner, John. "Intergenerational and Inter-household Economic Links," in Mark R. Rosenzweig and Oded Stark, eds., *Handbook of population and family economics*. Amsterdam: North-Holland, 1997, pp. 189-238.
- Mauss, Marcel. *The gift*. London: Routledge and Kegan Paul, 1966.
- Stark, Oded. *Altruism and beyond, An economic analysis of transfers and exchanges within families and groups*. Cambridge: Cambridge University Press, 1995.

The Prestige Motive for Making Charitable Transfers

By WILLIAM T. HARBAUGH*

If people are so self-interested, why do they give their money away to charities? One possibility is that people care about the level of the public good their donations provide. But this is not a good explanation: free-riding typically dominates donating even for people who care a great deal about the good in question, and even in groups that are substantially smaller than those in which people contribute.

An alternative explanation for giving is that the benefit comes from the donation itself, not from the good it buys. This idea is ancient. In the Old Testament, God promises those giving to the temple that he will

... open to you the windows of heaven,
and pour you out a blessing, that there
shall not be room enough to receive it.

(King James Bible, Malachi 3:10). More recently, Gary Becker (1974) has developed an economic model where it is the amount the donor gives, rather than the quantity of the public good he receives, that enters the utility function. James Andreoni (1989) shows that such a model can explain many of the observed facts of charitable giving, such as broad participation by people of different incomes, better than a model without this motive.

In this paper I consider two separate types of benefits that might arise from donations: those that are purely internal, derived from the donor's own knowledge of what he has given, and those that the donor only gets when other people know how much he has given. I call these the "intrinsic benefit" and the "prestige benefit," respectively.

Abundant descriptive evidence suggests that the prestige benefits from public recognition of donations are an important reason why people give. Large anonymous donations are so rare that they are newsworthy events. On the other hand, every university has buildings prominently named after alumni who gave substantial amounts, often only after the explicit promise of this sort of recognition. Most universities have a set price for those wishing to have a chair named after them. The prestige motive is important enough that the form of recognition the charity will provide in exchange for the gift is often spelled out in legal contracts, and there are even cases where donors have demanded the return of donations after their gifts have not been recognized to their satisfaction. For example, in 1997 a \$3 million donation to the New York City Children's Zoo was revoked by the donors after they argued that the city had not followed the contract which stipulated how their gift would be publicly recognized (David Dunlop, 1997).

In this paper I use data on gifts by lawyers to their law school to estimate a utility function which includes both prestige and intrinsic benefits. I estimate the parameters of this function by exploiting a common way by which charities report donations: publicizing the categories into which the donations fall, rather than the exact amounts. For example, the charity might say that donations of from \$500 to \$999 place the donor in the "Sustaining Contributor" category. Within this category, any portion of a donation that is above the \$500 lower bracket is not reported by the charity, and so provides no additional prestige, only additional intrinsic benefits. The more money people donate, the higher the tastes for both intrinsic and prestige benefits, while the greater the proportion of these donations that are just equal to the bracket, the higher is the taste for prestige.

To get a simple measure of the importance of the prestige effect, I will then use the estimated utility function to compare predicted

* Department of Economics, University of Oregon, Eugene, OR 97403-1285. I acknowledge the generous assistance of Oded Stark and useful comments from Zvi Griliches, Rick Harbaugh, and Louis Kaplow. Any remaining errors are my responsibility. By agreement with the institution providing the data used in this paper, they are confidential.

donations to the charity under two hypothetical plans: no reporting and exact reporting. A comparison of these results allows a measure of the effect of the prestige motive on donations. The distinction between the prestige and intrinsic motives for giving provides an interesting insight into behavior and is also important to charities and to society at large. While intrinsic benefits are obtained through the act of giving, and are therefore largely outside a charity's control, prestige is acquired only when a charity actually makes a public report of the amount of the donation. Common sense suggests, and virtually every "how to" book on fundraising agrees, that the actions of charities to solicit gifts and reward donors with public recognition have a large effect on giving and the voluntary provision of public goods. Despite this importance, there is currently little understanding of the workings of these factors, and no measure of the share of giving attributable to them. This paper will provide a measure of one part of that connection between the actions of the charity and donations, the part that is based on prestige.

I. Prestige and Intrinsic Benefits under Different Reporting Plans

Since a more complete model of the donor's optimization problem is given in Harbaugh (1998), I will only review the essential points here. I assume that donors have a utility function $U = U(x, p, d)$, where x is the private good, p is prestige, and d is the intrinsic benefit, assumed to be equal to the actual amount donated. The donor faces the budget constraint $w = x + qd$, where w is income and q is the after-tax price of giving. I assume that prestige is equal to the publicly reported amount of the donation and, for simplicity, that $q = 1$. Substituting the budget constraint into the utility function gives $U = U(w - d, p, d)$ or $U = V(p, d; w)$. Solving this for a fixed w and fixed levels of utility gives level curves in (d, p) space, as shown in Figure 1. Note that by construction the budget constraint must be satisfied along these curves, and that higher curves represent higher utility.

The reporting plans of the charity, perhaps most intuitively interpreted as an additional constraint, can be shown in the same (d, p)

space. These plans translate a donation d into a report r , which society then converts to the prestige p that enters the utility function. Three possible reporting plans are shown in Figure 1, along with level curves for a single donor with given wealth.

In the first plan, no reports are made, and the prestige function is a horizontal line at zero. The best donors can do is to give d_0 . In the second, charities report the exact amount of the donation, so the prestige function is the line $p = d$. The utility-maximizing donation is now d_e . Since a dollar donated now buys prestige as well as intrinsic benefits, donations can be expected to increase, unless the prestige reduces Ud or increases Ux significantly. In the third plan, the charity sets a category with a minimum amount or lower bracket needed to gain classification into that category. (I will examine situations where the charity sets more than one such category.) Those donating less than the amount of the lower bracket of the category get zero prestige, while those donating the bracket amount or more get credit for the amount of the category, as shown by the step function in Figure 1. Under this plan the optimal donation depends on preferences and the bracket. A person with preferences as shown will give the bracket amount d_b (or more) unless the bracket is above d_m . Note that under a given category reporting scheme a person's optimal donation may be either greater or less than what it would have been under exact reporting.

II. Data

I use exact and publicly reported donations by the class of 1976 to the alumni fund of a prestigious law school (the name of which I have agreed to keep confidential) for each year from 1989 to 1993. I also have income data for the alumni, obtained from a survey conducted in 1991. There are 223 complete observations, out of a class of 379. For estimation, I use only the 146 alumni who made at least one positive donation during this period.

The category reporting plan was changed in 1992. Prior to 1992, the lower brackets of the categories were \$100, \$250, \$500, and \$1,000. For 1992 through 1994, categories were \$500,

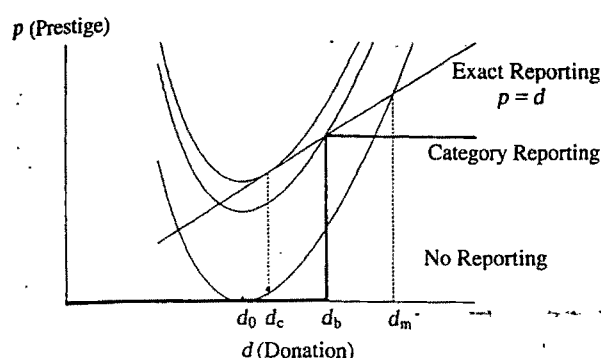


FIGURE 1. LEVEL CURVES AND REPORTING PLANS

\$1,000, \$2,500, and \$5,000. (There were also higher categories, in both cases. Only two donors gave at these levels, and they are excluded from the analysis.) This reporting change affected donations as the model predicts: the number of donations at the omitted categories fell. I use the reported 1991 income as the proxy for income in each year. This is a crude measure, and it might well be preferable to incorporate other information, such as the stock-market level, as a measure of year-to-year changes. I also assume that the after-tax price of giving is 0.66, for all donors and years. This is a serious approximation, in part because of individual variation in marginal tax rates, in part because of tax-law changes during this period, and in part because some employers, particularly law firms, match employee donations. (Information on which donors could take advantage of this is not available.) While it would be possible to estimate more appropriate prices, both this and the better income measures discussed above would complicate the econometric procedure substantially, since each donor in each year would then face a different budget constraint. As the measure of prestige associated with a given category, I use the average of all donations for the given year that fall within the brackets of that category.

III. Econometric Method

The econometric method I adopt is similar to methods developed for nonconvex budget constraints. Stephen Pudney (1989) describes these in detail. Since these methods involve determining the optimal donation by compar-

ing the utility from donations on each different portion of the constraint, a specific functional form for preferences is necessary. I use the Stone-Geary utility function:

$$U = \log x + b \log(p + k_1) + c \log(d + k_2)$$

where, as above, x is the private good, p is prestige, d is the intrinsic benefit (assumed to be equal to the donation), and k_1 and k_2 are constants. Donors maximize this function subject to the constraints of income, the price of donations, and the category reporting plan, which relates donations to p .

This function requires relatively few parameters. While it allows donations to be linear functions of income, which is convenient for the derivation of the likelihood function, it does not restrict donations to constant percentages of income, as the Cobb-Douglas utility function would. If k_1 is positive, it allows for the possibility of no reported donations (therefore no prestige). If k_2 is positive, it allows for the possibility of no donations (and therefore no intrinsic benefit). However, this functional form does impose weak separability and a linear expenditure system of demands.

One implication of this model of behavior is that donations just below the kink in the reporting function have zero probability. The maximum-likelihood method will therefore produce indifference curves steep enough to fit the donation closest to the kink, even if such curves do not fit the rest of the data well. One explanation for these observations is that they truly represent optimal behavior and should be used as they are in the estimation. Another, which I adopt, is that the donors can make mistakes. Donors will be uncertain about their preferences, about aspects of their budget constraint such as income or the tax treatment of a donation, and about the way in which the charity will report donations. I use two different models for errors. In the first, I simply assume that $\hat{d} = d^* + \varepsilon$, where \hat{d} is the actual donation, d^* is the optimal one, and ε is distributed $N(0, \sigma_\varepsilon)$. In the second, I attempt to account for the fact that donations just equal to the bracket will be optimal for a wide variety of preferences and incomes, and that therefore it is less likely that donors with optimal donations at a bracket will make errors

TABLE 1—ESTIMATES OF PREFERENCES

Error model	Estimated preferences		
1	$\ln x + 0.000276 \ln[p + 5.44]$ (0.00000449) (8.26) $+ 0.000256 \ln[d - 12.7]$ (0.00000676) (18.9)		
2	$\ln x + 0.000144 \ln[p - 49.9]$ (0.0000456) (4.04) $+ 0.000403 \ln[d + 29.6]$ (0.000127) (2.35)		
Error model	Log likelihood	σ_e	ν
1	-3,859	330 (4.26)	NA
2	-3,360	364 (5.70)	0.714 (0.0218)

Note: Standard errors are reported in parentheses.

that throw them off a bracket. I do this by assuming that alumni with optimal donations at a bracket will deviate from that bracket only with probability $\nu < 1$. If they do make an error, it is distributed as above, as are donations from donors whose optimal donation is not at a bracket. This model has an advantage over the first specification in that it tends to attribute errors to those for whom uncertainty about income and preferences should be most likely to lead them to change their donations, namely, those whose optimal donations are not at the brackets.

I find the parameters that maximize the probability of drawing the observed donations, given observed incomes. To find the probability of a given observation, I first take the utility-function parameters under consideration and the person's income and find his optimal donation, d^* . This is done by calculating the utility this person would get by making his optimal donation on each segment of the reporting plan and then picking the segment, and then the donation, where utility is highest. For the first error model, I then evaluate the probability density function for ε , $f(\varepsilon)$, at $\varepsilon = \hat{d} - d^*$, using the value of σ_e under consideration as the estimate of the error variance. The second model is handled analogously.

TABLE 2—PREDICTED DONATIONS UNDER DIFFERENT REPORTING SCHEMES

Income (\$)	Reporting scheme	Error model	
		1	2
50,000	none	32	1
	exact	45	65
150,000	none	70	61
	exact	123	127
300,000	none	127	151
	exact	242	243
Total (all observations)	none	119,100	99,001
	exact	151,589	134,744

Note: Donations are reported in dollars.

IV. Results

Table 1 gives parameter estimates and standard errors from the two models. Table 2 gives predicted donations under no reporting and under exact reporting for a variety of incomes and also for the entire data set. The income-specific predictions do not include the impact of the error term. Since donations are truncated at zero, and donations under exact reporting are always higher than under zero reporting, the effect of the truncation is to increase the expected donation under zero reporting by more than that under exact reporting, so it can be argued that these predictions overestimate the impact of reporting on donations. The predictions of overall giving (based on the income of the sample) in the last columns of the table do include the impact of these errors.

For brevity I only give information about fit for the first error model, results for the second are in general slightly better. The model overpredicts the numbers of small-bracket donations and underpredicts for larger brackets. The model prediction is 254 \$100-bracket donations, out of 720 total donations. As there are 115 actual \$100-bracket donations, chance would imply that 41 \$100 donations are correctly predicted as such, while the model correctly predicts 73. The model predicts 54 versus 32 actual \$250 donations, so chance would imply that two of these are correctly predicted as such, while the model correctly predicts eight. As there are very few donations at higher amounts, I omit discussion of those here.

I use the difference between donations under no reporting and under exact reporting as a simple measure of the incremental effect of the prestige motive, because this measure is not dependent on the particular brackets used under category reporting. The last two columns of the bottom panel of Table 2 show that under exact reporting donations would be from one-quarter to one-third above what they would be under no reporting, where they only depend on the intrinsic motive. If donors were not making optimization errors, the effect of the prestige motive would be stronger yet: many donors would double or even more than double their donations in response to the prestige motive. These amounts seem sufficient to warrant the importance charities attach to reporting gifts and recognizing donors, and they corroborate the descriptive evidence that, at least for some donors, the prestige motive is quite strong.

V. Conclusions

By examining category reporting this paper has concentrated on the simplest and most accurately measurable way by which the actions of a charity can, through a prestige effect, influence the amount donors give. The results support the hypothesis that donors have a taste for prestige, and they show that a substantial portion of donations can be attributed to it.

Some caveats should be made. First, there is clearly a substantial amount of heterogeneity among donors, and it might be more appropriate to use a model that explicitly allows for such heterogeneity. Additionally, donors may give bracket amounts simply because such amounts are focal points noted on the mailing envelope, rather than because the donors are explicitly optimizing in the face of intrinsic and prestige benefits. This effect will bias my estimate of the prestige motive upward.

These are atypical donors, giving to an atypical charity. Lawyers have good reasons to signal that they are successful. Donations to law schools are an obvious way to do this, so the prestige motive may be stronger for these donations than for those to, say, the Salvation Army. On the other hand, there are also reasons the prestige motive might be weaker for

these donors than is typical. Relative to other groups who make public donations, a law school class is small, and members have many ways to gain information about each other outside of seeing names and donor categories in the alumni magazine. The prestige motive might be considerably stronger in other cases, say, for a public figure who wants people who do not know him personally to be aware of his generous behavior.

Perhaps the most important and potentially interesting caveat is that there are alternative ways to model how donations, and public knowledge about donations, enter utility. For example, both the intrinsic and the prestige benefits from a donation may not simply be equal to the dollar amount, but instead might be relative to gifts by others. If so, it seems likely that a person's donations are mainly compared with those of a "reference group," as in Oded Stark (1990), composed of people the donor knows and with whom he shares common experiences and characteristics.

The importance of reference groups might explain why fund-raisers often emphasize such social activities as parties, dinners, and reunions: these strengthen such groups. The importance of relative donations within these groups may explain the common practice of having large donors solicit contributions from others in their circle. People should presumably increase their donations after being told that a member of their group has given a large amount, especially after they have just had dinner with him.

In this paper I have used category reporting and the tendency for donors to give amounts equal to the lower brackets of categories only as means of estimating the importance of the prestige and intrinsic motives, given my assumptions about how these enter the utility function. In Harbaugh (1998) I examine the characteristics of the donation-maximizing category plan for a charity that has donors with these same sorts of preferences. In future work I plan to combine these approaches in an attempt to learn more about precisely how the intrinsic and the prestige benefits affect the utility of donors. For example, if the prestige benefit depends on how a person's reported donation compares to donations by others, the charity's optimal reporting plan will

presumably be different than if prestige only depends on the amount itself. Assuming that fund-raisers know their donors and want to maximize donations, it should be possible to make inferences about donors' preferences from the actual reporting plans.

REFERENCES

- Andreoni, James.** "Giving with Impure Altruism: Applications to Charity and Ricardian Equivalence." *Journal of Political Economy*, December 1989, 97(6), pp. 1447-58.
- Becker, Gary.** "A Theory of Social Interactions." *Journal of Political Economy*, November-December 1974, 82(6), pp. 1063-94.
- Dunlap, David W.** "\$3 Million Zoo Gift Revoked Because Plaque Is Too Small." *New York Times*, 15 May 1997, Section B, p. 1.
- Harbaugh, William T.** "What Do Donations Buy? A Model of Philanthropy Based on Prestige and Warm Glow." *Journal of Public Economics*, February 1998, 67(2), pp. 269-84.
- Pudney, Stephen.** *Modelling individual choice: The econometrics of corners, kinks and holes*. Cambridge, MA: Blackwell, 1989.
- Stark, Oded.** "A Relative Deprivation Approach to Performance Incentives in Career Games and Other Contests." *Kyklos*, 1990, 43(2), pp. 211-27.

Tax Policy and Gifts

By LOUIS KAPLOW *

Voluntary transfers between individuals are potentially subject to income taxes and wealth transfer (estate and gift) taxes. With regard to the income tax, Henry Simons (1938) argued that it should be levied both on the donor, whose gift is a form of personal consumption, and on the donee, who directly consumes the gift. Others would limit income taxation to the donee, the only one whose act of consumption dissipates real resources. Most income tax systems do tax gifts only once, but the single tax is imposed in a different, more administratively convenient manner: the donee is exempt, and instead the donor is taxed, implicitly, by not allowing any deduction for gifts. Yet the rationale for simply applying the labor income tax rate—whether once or twice—to gifts is dubious because, as I will discuss, the incentive, distributive, and other welfare effects of taxing gifts and of taxing labor income are different.

Wealth transfer taxes, which in most developed countries are levied only on the estates of wealthy individuals, are often evaluated in terms of their redistributive effects. But such analysis is not usually integrated with that of the income tax, which is also a redistributive tool. Another factor suggesting the need for a more integrated treatment of transfer taxes and income taxation is that the effects of taxation on behavior and welfare will depend on the aggregate of taxes levied on a gift rather than on what portion of the tax is designated as a gift or estate tax and what portion is deemed to be an aspect of the income tax.

Accordingly, this paper considers a single, unified framework for analyzing the combined taxation of gifts, one that incorporates existing

analysis of redistributive income taxation.¹ Using such an approach, I sketch a mapping between types of gifts and optimal tax policy. It is helpful, however, to begin by discussing how gifts should be taxed if they were simply another form of ordinary consumption.

I. Gifts as a Form of Ordinary Consumption

A. Separability of Redistribution

The approach here is to begin with the standard optimal labor income tax problem in a world with no gifts and then to examine how that tax should be altered in the presence of gifts. Interestingly, under such a formulation of the transfer taxation question, redistribution becomes largely a separate issue. The reason is that redistribution is accomplished directly, by adjusting the tax schedule as a function of income. The transfer taxation problem involves determining whether, say, parents who give an above-average fraction of their income to their children should be taxed more or less *relative to other parents at the same income level* who instead spend a greater fraction of their income on themselves.

This preliminary study will not formally model this rather complicated optimal income tax problem. To gain some initial insights, it is useful to undertake a simpler thought experiment. Suppose that, at each level of income for donors, more (less) generous treatment of gifts is achieved by lowering (raising) the tax rate applicable to income expended on gifts and raising (lowering) the tax

¹ This inquiry abstracts from such questions as whether lifetime gifts and bequests should be distinguished; how the substantial majority of wealth, human capital (created in large part through different sorts of transfers), should be incorporated; what the effects are on savings and whether they may be ignored because they can be offset by other government policies; and what additional issues are raised by transfers to charities.

* Harvard Law School, Cambridge, MA 02138, and NBER. I am grateful to Steven Shavell, Oded Stark, Alvin Warren, and seminar participants at Harvard University and the NBER for comments and to the John M. Olin Center for Law, Economics, and Business at Harvard Law School for financial support.

rate applicable to income expended on direct consumption, that is, labor income net of any gifts made. For example, a gift subsidy could be understood as a tax credit for gifts combined with a higher labor income tax rate, so that the tax on donors' direct consumption would be higher and the net burden on gifts would be lower than under a uniform system. One can thus hold the total tax burden on each income class constant. (Although this may not be optimal, the approach focuses attention on the treatment of gifts relative to that of direct consumption.)

B. *Optimal Relative Taxation of Gifts and Ordinary Consumption*

As a benchmark, it is efficient to tax individuals' expenditures on different goods and services in the same manner (i.e., there would not be differential commodity taxation) when there is an income tax. The standard qualification is that taxing more heavily (lightly) expenditures on commodities that make leisure relatively more (less) attractive would lessen the labor/leisure distortion caused by income taxation (see A. B. Atkinson and Joseph Stiglitz, 1976).² Thus, viewing gifts for the moment as simply another form of consumption, it is efficient to tax them relatively more heavily if, say, donors need more leisure time to enjoy their utility from giving (such as by spending time with the children whom they support) than to enjoy other forms of consumption. On the other hand, it is efficient to tax gifts more lightly if, for example, much of the enjoyment is vicarious, deriving from contemplation of the gift. With bequests in particular, a donor may work harder in order to leave a larger bequest, whereas workers who instead spend their earnings on vacations would need more leisure time. The question of whether it is efficient to tax or to subsidize gifts, viewed as another form of ordinary consumption, is an empirical one that has not, to my knowledge, been investigated.

C. *How Gifts May Differ*

The remainder of this paper will focus on the manner in which gifts differ fundamentally

from ordinary personal consumption: making a gift does not expend real resources, but instead shifts them to another individual.³ I will now explore the implications of this difference and how they depend on the type of gift that is involved.

II. Altruism

A. *Positive Externality on Donees*

Gifts convey a sort of positive externality on donees.⁴ To see this, consider the case of an altruistic donor who equally values the direct utility from his own consumption and the donee's utility from her own consumption:

$$U(x, y, g) = u(x - g) + v(y + g)$$

where g is the amount of the donor's gift, x and y are the donor's and donee's pretransfer incomes, and $u(\cdot)$ and $v(\cdot)$, assumed to be strictly concave, are the donor's and donee's utilities from their own consumption.⁵ Observe that the donor counts the benefit to the donee as it enters the donor's own utility function, whereas a social welfare assessment should also weigh in the benefit to the donee. Thus, under a utilitarian social welfare function, $W(x, y, g) = u(x - g) + 2v(y + g)$. This discrepancy between the donor's and society's objectives suggests that treating gifts more generously than own consumption and, hence, a gift subsidy might be optimal.

To explore this point, consider the following formulation of the donor's utility function:

$$\begin{aligned} U(x, y, g) \\ = \alpha u(x - (1 - s)g - t) + \beta v(y + g) \end{aligned}$$

where α and β are the weights that the donor gives to the direct utility from his own con-

³ It is also relevant that gifts may involve a sort of voluntary redistribution, such as in the case of intergenerational transfers when there is regression toward the mean in earnings ability (see e.g., D. L. Bevan and Stiglitz, 1979).

⁴ This idea has been noted by Atkinson (1971), among others, and is developed in Kaplow (1995).

⁵ This is a special case of the model introduced in Gary Becker (1974).

² Other qualifications will not be explored here.

sumption and to the donee's utility from her own consumption, s is a subsidy on gifts, and t is a tax (taken as given by the donor) that the government collects in order to finance the subsidy.

Now, consider the effects of marginally increasing s , beginning at $s = 0$. First, this will induce the donor to increase his gift. This can be seen from the donor's first-order condition:

$$\alpha(1-s)u'(x - (1-s)g - t) \\ = \beta v'(y + g).$$

Raising s (and also increasing t to finance the increase in s) will, at a given level of g , reduce the value of the left side; to restore equality, g must increase, given the strict concavity of $u(\cdot)$ and $v(\cdot)$.⁶ As a consequence, the donee's utility will increase. Moreover, at $s = 0$, it can be shown that there will be no first-order effect on the donor's utility.⁷ Finally, the donor's net position vis-à-vis the government also is unchanged because the increase in the subsidy on gifts is financed by taxing donors. In sum, a slight increase in the subsidy will help the donee at no cost to the donor or the treasury.

B. Externality with Respect to Labor Income Tax Revenue

When considering possible tax-revenue effects in a world with a labor income tax (which is not explicitly modeled here), it is appropriate to take into account possible changes in labor supply. For donors, a small change in s , beginning at $s = 0$, will have no direct labor-supply effect: because the donor's

utility remains the same for any given level of earned income, the choice of labor effort would be unaffected.⁸ (There is, of course, the qualification noted previously for the case in which changing the donor's allocation of income between gifts and own consumption changes the relative value of leisure.)

Gifts may, however, result in a tax-revenue externality because they augment donees' income. In conventional analyses, this externality would be negative: donees would work less because of the income effect and thus would pay less income tax. To combat this externality, one could tax donees on the gifts they receive (or, equivalently, tax donors' gifts more heavily).

There are, however, other considerations. Gifts to donees might relax liquidity constraints that otherwise limit investments in human capital or entrepreneurship (see Douglas Holtz-Eakin et al., 1994). The net long-run effect of such gifts may be to increase donees' earnings and thus donees' tax payments, creating a positive tax-revenue externality. And there may be strategic effects that would influence donees' earnings and, thereby, the income taxes they pay: donees might choose to earn less, knowing that their plight will induce altruistic donors to give more (the "Samartitan's dilemma"), or donees might undertake activities that increase their income because donors might promise future gifts that are conditional on such behavior.

C. Gifts' Effects on Donors' and Donees' Marginal Utility

Many models of altruism assume that

$$U(x, y, g) = u(x - g) + \beta v(y + g).$$

That is, $\alpha = 1$ and $\beta > 0$. First, consider how the marginal utility of consumption of donors compares to that of nondonors. Donors, as a consequence of their giving, would have lower

⁶ Differentiating the first-order condition with respect to s , where $t = gs$, yields

$$\frac{dg}{ds} = \frac{-\alpha u'}{\alpha(1-s)u'' + \beta v''} > 0.$$

⁷ The derivative is

$$\frac{dU}{ds} = g'(-\alpha(1-s)u' + \beta v' - \alpha s u').$$

On the right side, the first two terms in parentheses, taken together, equal zero (from the donor's first-order condition), and the third term is zero at $s = 0$.

⁸ At $s > 0$, the induced increase in g reduces the donor's own consumption and thereby raises his marginal utility of consumption; ceteris paribus, this would tend to increase labor effort.

own consumption than would nondonors at the same income level. As a result, donors would have a higher marginal utility of income than would nondonors who had identical functions $u(\cdot)$ for utility of own consumption. (This case might arise, for example, when donors differ from nondonors only in that the former are fortunate enough to have found compatible mates or to have had children who please them sufficiently to induce giving.) In maximizing a utilitarian social welfare function, this difference in marginal utilities would warrant more favorable tax treatment of donors.⁹

Now suppose instead that donors are not individuals who place an unusually high value on others' well-being, but instead are those who derive unusually low utility from their own consumption; that is, their α is much less than 1 (even though, perhaps, still greater than β). Then, in spite of donors' lower consumption, their marginal utility might be lower than that of nondonors whose utility from own consumption was given by $u(\cdot)$, which would justify less favorable treatment under a utilitarian social welfare function.

Distinguishing these two cases requires making interpersonal utility comparisons. (A donor's observable behavior depends only on α/β and not on the absolute magnitude of α and β .) Analysts often elude this problem by stipulating that all individuals have the same utility function, but this assumption cannot be maintained in the present context because it is inconsistent with the heterogeneity in the behavior that is under consideration—namely, some individuals are donors and others are not. Hence, judgments about transfer policy, like judgments about general redistribution policy, must to some extent reflect views about different individuals' utility functions that cannot be grounded in observable behavior.

Another factor is that donees' own consumption will be higher than otherwise on account of the gifts that they receive. This

implies that their marginal utility of income will be lower than that of individuals with the same earned income who do not receive gifts, which favors heavier taxation of donees (tantamount to heavier taxation of gifts). Nevertheless, it remains true that, abstracting from possible tax-revenue externalities, it will be optimal for objects of altruism (prospective donees) to receive more effective income and thus have a lower marginal utility of income than others because their utility from own consumption receives additional weight in the social welfare function due to altruists' concerns.

III. Utility from Giving Per Se

Some donors may care only about the gifts that they themselves make, motivated by a need for self-sacrifice or a desire for prestige, as in James Andreoni (1990). Let

$$U(x, g) = u(x - (1 - s)g - t, g).$$

Analysis of this case, it turns out, is virtually the same as for altruism. It can be shown that introducing a positive subsidy will induce the donor to give a larger gross gift, and that this will increase the utility of the donee without reducing the donor's utility (at $s = 0$).¹⁰ Tax-revenue effects and other factors will be analogous as well.

The preceding formulation may, however, be inappropriate. If the donor really is moti-

¹⁰ The donor's first-order condition is now

$$(1 - s)u_1 = u_2$$

where a subscript i denotes the derivative with respect to the i th argument. Differentiating this condition yields

$$\frac{dg}{ds} = \frac{-u_1}{(1 - s)u_{11} - (1 - s)u_{12} - u_{21} + u_{22}}.$$

The numerator is negative, and from the second-order condition, the denominator is negative, so $dg/ds > 0$. Finally,

$$\frac{dU}{ds} = g'(-(1 - s)u_1 + u_2 - su_1).$$

Again, the first two terms in parentheses on the right side, taken together, equal zero (from the donor's first-order condition), and the third term is zero at $s = 0$.

⁹ Different social welfare functions may have qualitatively different implications. For example, under a maximin function, such altruistic donors are better off than others on account of their altruistic preferences and thus should be taxed more heavily (as should donees, who are better off on account of receiving altruists' gifts).

vated in a manner that depends on his own sacrifice and not on the gross gain to the donee from all sources (as with the altruist), then it seems reasonable that the donor's benefit should not depend on the gross gift, g , as in the preceding model, but rather on the net amount that he himself gives up, $(1 - s)g$.¹¹ Accordingly, consider

$$U(x, g) = u(x - (1 - s)g - t, (1 - s)g).$$

In this case, the effect of a gift subsidy on social welfare is quite different: raising s directly reduces the donor's utility because he benefits only from his own sacrifice, $(1 - s)g$, which is reduced as s is increased. It turns out that when a subsidy induces the donor to increase his gift, he is in essence redistributing his own income to the donee; unlike the previous cases, here a higher gross gift induced by a subsidy produces a utility benefit to the donee but a utility loss to the donor (even at $s = 0$).¹² The limited empirical work on this transfer motive has not sought to distinguish between these two different formulations.

IV. Exchange

If donors' gifts are in exchange for donees' efforts (see Douglas Bernheim et al., 1985;

¹¹ To dramatize the point further, suppose that the subsidy was not paid to donors, but instead was administered in the financially equivalent form of a matching grant paid directly to donees. (Note that an assumption implicit in both formulations is that donors do not derive utility from paying taxes, even if those taxes are used to subsidize gifts.)

¹² The donor's first-order condition is now $u_1 = u_2$. From this, one can derive:

$$\frac{dg}{ds} = \frac{-gu_{12} + gu_{22}}{u_{11} - (1 - s)u_{12} - u_{21} + (1 - s)u_{22}}.$$

Observe that dg/ds may not be positive. (It will be unless u_{12} is sufficiently negative.) Finally,

$$\frac{dU}{ds} = g'(-u_1 + u_2 - su_2) - gu_2.$$

Again, the first two terms in parentheses on the right side, taken together, equal zero (from the donor's first-order condition), and the third term is zero at $s = 0$. But now there is the additional term, which is negative even at $s = 0$.

Donald Cox, 1987), the transaction really consists of ordinary consumption by the donor and labor income earned by the donee, and each component should be taxed accordingly. A different form of exchange arises when transfers actually are loans (or loan repayments) or elements of various forms of insurance and annuity schemes (see Laurence Kotlikoff and Avia Spivak, 1981). In such cases, payments in both directions would generally be exempt from taxation.

Other posited transfer behavior has elements of exchange. Oded Stark and Ita Falk (1998) suggest that individuals may make gifts to engender gratitude in recipients, who may later return the favor. In this case, one of the preceding types of exchange may be present, depending upon whether the initial gifts or later "repayments" comprise labor effort. James Buchanan (1983) claims that potential donees will engage in rent-seeking behavior to elicit gifts from prospective donors, in which case perhaps gifts should be taxed, if donees' efforts to induce gift-giving waste resources.

V. Conclusion

This paper offers a framework for assessing tax policy with regard to private voluntary transfers to individuals. The main elements are integrating the income tax and estate/gift tax's treatment of gifts, taking a unified view of the distributive problem in the context of an optimal income tax framework, and focusing on aspects of gifts that distinguish them from donors' ordinary consumption.

The present analysis reveals that the optimal tax treatment of gifts (relative to the tax treatment applicable to labor income that is expended on direct consumption for oneself) is extremely sensitive to the type of gift involved:

- (i) *Altruism*.—Gifts involve a positive externality on donees, which favors a subsidy; there may also be a positive or negative tax-revenue externality, and donors may have higher or lower marginal utility than others, which tends to favor a larger subsidy or a smaller subsidy (or a tax), as the case may be.
- (ii) *Utility from giving per se*.—This is

similar to altruism if the donor's utility depends on the gross gift (i.e., including the subsidy), but there is no positive gift externality if the donor's utility depends on the gift net of the subsidy.

- (iii) *Exchange*.—If a gift is really compensation in exchange for labor, the "gift" should be taxed as part of labor income; if the exchange is financial (such as with loans and repayments or insurance arrangements), no tax or subsidy is appropriate; if gifts are induced by wasteful rent-seeking, a tax may be optimal.¹³

These results indicate the policy relevance of further empirical work that distinguishes among transfer motives and identifies more precisely the form of donors' utility functions. Because transfer motives no doubt vary greatly among donors, and in ways that the government cannot readily observe, it may be necessary to adopt tax policies based upon average behavior or to employ some simple categorical rules that, perhaps, distinguish among gifts between spouses, transfers to descendants, and contributions to public charities, based upon the typical characteristics of each class of gifts.¹⁴

REFERENCES

- Andreoni, James.** "Impure Altruism and Donations to Public Goods: A Theory of Warm-Glow Giving?" *Economic Journal*, June 1990, 100(401), pp. 464–77.
- Atkinson, A. B.** "Capital Taxes, the Redistribution of Wealth and Individual Savings."

Review of Economic Studies, April 1971, 38(2), pp. 209–27.

- Atkinson, A. B. and Stiglitz, Joseph E.** "The Design of Tax Structure: Direct versus Indirect Taxation." *Journal of Public Economics*, July–August 1976, 6(1–2), pp. 55–75.

- Becker, Gary S.** "A Theory of Social Interactions." *Journal of Political Economy*, November–December 1974, 82(6), pp. 1063–93.

- Bernheim, B. Douglas; Shleifer, Andrei and Summers, Lawrence H.** "The Strategic Bequest Motive." *Journal of Political Economy*, December 1985, 93(6), pp. 1045–76.

- Bevan, D. L. and Stiglitz, J. E.** "Intergenerational Transfers and Inequality." *Greek Economic Review*, August 1979, 1(1), pp. 8–26.

- Buchanan, James M.** "Rent Seeking, Noncompensated Transfers, and Laws of Succession." *Journal of Law and Economics*, April 1983, 26(1), pp. 71–85.

- Cox, Donald.** "Motives for Private Income Transfers." *Journal of Political Economy*, June 1987, 95(3), pp. 508–46.

- Holtz-Eakin, Douglas; Joulfaian, David and Rosen, Harvey S.** "Entrepreneurial Decisions and Liquidity Constraints." *Rand Journal of Economics*, Summer 1994, 25(2), pp. 334–47.

- Kaplow, Louis.** "A Note on Subsidizing Gifts." *Journal of Public Economics*, November 1995, 58(3), pp. 469–77.

- . "Optimal Distribution and the Family." *Scandinavian Journal of Economics*, March 1996, 98(1), pp. 75–92.

- Kotlikoff, Laurence J. and Spivak, Avia.** "The Family as an Incomplete Annuities Market." *Journal of Political Economy*, April 1981, 89(2), pp. 372–91.

- Simons, Henry C.** *Personal income taxation*. Chicago: University of Chicago Press, 1938.

- Stark, Oded and Falk, Ita.** "Transfers, Empathy Formation, and Reverse Transfers." *American Economic Review*, May 1998 (*Papers and Proceedings*), 88(2), pp. 271–76.

¹³ Note also that purely accidental bequests (when individuals cannot annuitize), which were not examined here, might optimally be subject to confiscatory taxation.

¹⁴ Such distinctions are made in current regimes (e.g., in rules defining the tax and welfare treatment of various family units), although existing rules are not well rationalized in terms of the general distributive objectives of the tax system or the motives likely to govern sharing (transfers) within the family. For an initial study of this problem, see Kaplow (1996).

TAX AND HUMAN-CAPITAL POLICY†

Taxes, Uncertainty, and Human Capital

By KENNETH L. JUDD*

Human capital is the most important determinant of wealth and income for most individuals. However, income-tax analysis has devoted far less effort to understanding the taxation of human capital than the taxation of physical capital and labor supply. A separate treatment is necessary since human capital is neither just capital nor just labor supply.

This essay begins by examining optimal taxation in a dynamic model with human capital. I then try to apply the insights from these models to tax policy, quickly finding that even a superficial treatment forces consideration of issues in political theory, contract theory, and financial theory. I take issue with some of the usual assumptions of the literature and show how alternative views affect the interpretation of optimal-tax rules. In particular, I argue that many so-called consumption-tax proposals are not true consumption taxes and are biased against human capital.

I. Optimal-Tax Results

I first review some results from an optimal-tax analysis. Education is an investment good since it increases wages, but it may also have consumption value. Peter Diamond and James Mirrlees (1971a, b) argue for no taxation of intermediate goods. Judd (1985) argues for no tax on capital in the long run for any Paretian objective and for Uzawa preferences, showing instead that only wage (or consumption) taxation is desirable in the long run. Since human capital is a mixture of labor supply, capital in-

vestment, and final good, the implications of these ideas for human capital are unclear.

Judd (1997a) examines these issues in a dynamic model. Specifically, the paper assumes that the representative individual solves the following problem:

$$\max_{c, n, g, x} \int_0^{\infty} e^{-\rho t} u(c, n, H, g) dt$$

subject to

$$\dot{A} = \bar{r}A + \bar{w}L(H, n) - c - x - \tau_H H$$

$$\dot{H} = x$$

where H is human capital, $L(H, n)$ is effective units of labor given n hours of labor and human capital H , A is financial assets, τ_H is a tax on human capital, \bar{r} is the after-tax return on financial assets, \bar{w} is the after-tax wage for a unit of effective labor, and x is human-capital investment. The production function is $f(k, L(H, n), g, t)$ where k is physical capital, g is government consumption, and the argument t models exogenous productivity trends. The government chooses \bar{r} , \bar{w} , and τ_H to maximize the representative agent's utility subject to the government's intertemporal budget constraint.

The incorporation of human capital in this problem generates a tension. If one thinks of human capital as capital then the logic in Judd (1985) argues for no taxation of H , leaving only labor-income taxation in the long run. However, it is difficult to tax labor income without distorting human-capital investments. Judd (1997a) shows that if $u_H = 0$ then there should be no long-run net taxation on returns to human-capital investment, only taxation of hours of labor supply. This can be implemented by taxing labor income but immediately expensing all human-capital investment expenditures. To do this, I set $\bar{w} < f_L$ and choose negative values for

† Discussant: Nancy L. Stokey, University of Chicago.

* Hoover Institution, Stanford, CA 94305, and the National Bureau of Economic Research. The author gratefully acknowledges the support of NSF grant SBR-9309613.

τ_H so that taxable income equals $Lf_2 - Hf_1$, which is labor income minus the opportunity cost of human capital.

If H is a final good (or bad), that is, $u_H \neq 0$, a positive tax on human-capital formation may be desirable. While it may seem odd to take seriously the notion of human capital being a consumption bad, this is the case if there are nonpecuniary and non-time costs associated with acquiring human capital. For example, some students find the cost of taking a calculus course to exceed the monetary value of the time spent in lecture and doing homework.

II. Tax Treatment of Human-Capital Investments

The tax rules derived above are clear in a simple model, but interpreting them for the real world is difficult. I will first consider the case where H is not a final good, implying that all human-capital expenditures should be expensed. The U.S. tax code takes a mixed approach. On-the-job training and a student's own time are both effectively expensed, while expenditures such as tuition and books are generally not deductible. It also appears that human capital is taxed less, since human capital is taxed only at the personal level whereas asset income is also taxed at the corporate level.

However, the picture is more complex. The typical analysis treats educational expenditures of state and local governments as subsidies. The Tiebout theory of local publicly provided goods argues otherwise. Local education expenditures are financed largely by local taxation and controlled largely by local political entities. The Tiebout view argues that public education expenditures are effectively equivalent to private expenditures. Combining Tiebout with the optimal-tax analysis, I conclude that local public education expenditures should be deducted from the tax base. The general point is that if citizens of a community decide to finance jointly the education of their children through local taxes and those expenditures respond to the after-tax cost, then a deduction is desirable.

This is currently implemented partially by the deductibility of state and local income and property taxes in the federal income tax. In

contrast, some parents pay substantial nondeductible tuition to send their children to private schools. This difference creates a bias in favor of public schools. Also, itemization is more common among high-income families, implying a regressive tax on human-capital accumulation. The optimal-tax results argue for the deductibility of all these expenditures in all communities.

The subsidy view of education is also challenged when one looks at allocation rules. Tuition at state universities is generally lower than cost, but that does not imply a net subsidy to individuals, since admission is often rationed. Since even rationed markets must "clear" there will be some other private costs to equate demand and supply. For example, a student may work harder during high school to get into a preferred university.

Contrary to common opinion, financial investments may be taxed less than human-capital investments. Pension funds, IRA's, and 401(k) plans allow workers to avoid personal taxation of their retirement savings. Investments in corporate debt also avoid the corporate income tax. Furthermore, owner-occupied housing is treated favorably. Finally, in a progressive tax system, the deduction of a student's time occurs when his tax rate is low but the return to human capital is greatest during working years when the tax rate is highest. In contrast, pension-fund contributions are deducted at working-year tax rates, but pension distributions are taxed during retirement years when the tax rate is likely to be lower.

These considerations are not minor. In fact, 1990 total expenditures on education (other than federal aid) was \$370 billion compared to \$576 billion in gross investment in nonresidential fixed capital. Human-capital tax issues are not small relative to intensely debated business-taxation issues.

The relative importance of these factors differs substantially across individuals. The combination of a Tiebout model and low taxation of financial assets often applies to a stereotypical upper middle-income family owning their home in a suburban community, but it is less plausible for a poor inner-city family renting an apartment. Also, one story may be true for most individuals, but another true for most wealth. Fortunately, the ideal tax system is

much clearer: no taxation of intermediate goods and services, and to the extent that education is an intermediate good its costs should be expensed.

III. Human Capital and Idiosyncratic Risk

The previous analysis has ignored risk. All investments are risky, but the risk to human-capital returns has extra dimensions. Investors can diversify across various firms in their financial portfolios, but human-capital diversification is difficult. Jonathan Eaton and Harvey Rosen (1980) have modeled wage uncertainty as idiosyncratic, uninsurable risk and argued for a subsidy to education. If true, then arguments for generous tax treatment of education would be strengthened. Since risk is an important feature of any investment, one needs to take this issue seriously in any analysis.

Unfortunately, any conclusion is sensitive to the kind of idiosyncratic risk. To see this, consider the simple model of human-capital accumulation with moral hazard examined in Judd (1997b). Suppose one can invest s in a safe asset with return R or h in human capital. Assume that an individual's output will be $f(h)$ with probability p and zero otherwise, where p is chosen by the worker and not observed by his employer. The employer pays a wage w_1 if the worker is successful and w_2 otherwise. An optimal (and competitive-equilibrium) contract with a risk-neutral employer then specifies h , s , p , and state-contingent wages to maximize the worker's expected utility so that his wage income equals his output in expectation, and so that the choice of p is incentive-compatible. This leads to the following problem:

$$\max_{w_1, w_2, s, h, p} E\{u(w + sR)\} - v(p)$$

subject to

$$pf(h) - T - E\{w\} = 0$$

$$u(c_1) - u(c_2) - v'(p) = 0$$

$$1 - s - h = 0$$

where $u(c)$ is utility over consumption and $v(p)$ is the disutility of effort.

A straightforward analysis shows that $pf'(h) - R = 0$ in equilibrium. Therefore, h is chosen to equate the expected marginal product of h with the rate of return on the safe asset, implying no risk premium. Furthermore, since the contract is constrained efficient, there is no efficiency argument for subsidization of human capital. This illustrates a more general point. If private agents have as much information as the government and can write flexible contracts, then it is unclear how relatively inflexible tax policies can dominate private contracts.

IV. Education and Tax Reform

The exercises above argue that the most important features for human capital are its productivity and final-good properties. Perhaps the idiosyncratic risk features are also important, but the results are not robust to alternative models of idiosyncratic risk. With this in mind, I next discuss some tax-policy proposals.

There are several proposals for radical income-tax reform which claim to be consumption taxes, the flat tax proposal of Robert Hall and Alvin Rabushka (1995) being a typical example. However, their view of investment is limited. If one defines consumption to be output minus investments, then the flat tax fails to be a consumption tax since it denies expensing of many human-capital investments. It would be worse than the current tax system for some, since it would eliminate the deductions for state and local taxes and charitable contributions.

Since taxation of physical capital would be eliminated, these so-called consumption-tax proposals would produce biases in favor of physical capital. This may be justified by concerns for simplicity or a belief that we should discourage state and local governments from spending money on education, but not by consumption-tax principles.

V. Is Education Only an Intermediate Good?

The above examples appear to come to a simple conclusion: all educational expenditures should be treated as investment, and all investment should be expensed in an ideal tax system. However, they rely on the common assumption, which I provisionally made

above, that human capital is only an intermediate good. I complete this essay by asking, "Is education only an intermediate good?"

This can be determined by comparing financial returns of alternative assets. If human capital has a lower financial return than financial assets of *comparable riskiness*, then human capital must have some nonpecuniary returns. As Gary Becker (1975) argues, education and corporate equity have roughly the same mean financial return. Why does education have as high a risk premium as equity? Unfortunately, there is little empirical work on this. Perhaps the premium is due to idiosyncratic risk. However, the moral-hazard example showed that there is no justified premium in some circumstances. Wage income may have a positive covariance with profits, but wages are less cyclical than profits. Furthermore, the price of risk for human capital depends on the covariance between profits and the *marginal* impact of human-capital investment on risky wages (see Judd, 1997b). Since the less-educated and less-experienced are the most likely to experience unemployment during a recession, education would appear to reduce one's exposure to systematic risk. Therefore, the price of risk to be attached to human-capital investments appears to be smaller than that associated with corporate equity.

If future empirical analysis confirms these impressions, then human capital has a mean return greater than comparable financial assets. In a perfectly competitive market, this implies that human capital is a consumption bad, and according to the optimal-tax analysis, that human capital should be taxed. Of course, there are other explanations for the risk premium which could be put forward, such as liquidity constraints, political inefficiency, and imperfect altruism, many of which imply underinvestment in education. These alternative explanations would have substantially different implications for tax policy.

VI. Conclusion

The theory of optimal taxation of human capital is rather simple, but it is difficult to apply it since there is so little empirical evidence about the critical determinants. In particular, there is a need for better understanding of local government decision-making and the riskiness of human-capital investment. I have no answers to these issues. These simple examples, though, show how asking simple questions about the taxation of human capital quickly leads to a variety of difficult, important, and unresolved problems.

REFERENCES

- Becker, Gary. *Human capital: A theoretical and empirical analysis*. Chicago: University of Chicago Press, 1975.
- Diamond, Peter A. and Mirrlees, James A. "Optimal Taxation and Public Production: I—Production Efficiency." *American Economic Review*, March 1971a, 61(1), pp. 8–27.
- . "Optimal Taxation and Public Production II: Tax Rules." *American Economic Review*, June 1971b, 61(3), pp. 261–78.
- Eaton, Jonathan and Rosen, Harvey S. "Taxation, Human Capital, and Uncertainty." *American Economic Review*, September 1980, 61(4), pp. 705–15.
- Hall, Robert E. and Rabushka, Alvin. *Low tax, simple tax, flat tax*, 2nd Ed. New York: McGraw-Hill, 1995.
- Judd, Kenneth L. "Redistributive Taxation in a Simple Perfect Foresight Model." *Journal of Public Economics*, October 1985, 28(1), pp. 59–83.
- . "Optimal Taxation and Spending in General Competitive Growth Models." Mimeo, Hoover Institution, 1997a.
- . "Is Education as Good as Gold?" Mimeo, Hoover Institution, 1997b.

Tax Policy and Human-Capital Formation

By JAMES J. HECKMAN, LANCE LOCHNER, AND CHRISTOPHER TABER *

Missing from recent discussions of tax reform is any systematic analysis of the effects of various tax proposals on skill formation (see the papers in the collection edited by Henry Aaron and William Gale [1996]). This gap in the literature in empirical public finance is due to the absence of any empirically based general-equilibrium models with both human-capital formation and physical-capital formation that are consistent with observations on modern labor markets. This paper is a progress report on our ongoing research on formulating and estimating dynamic general-equilibrium models with endogenous heterogeneous human-capital accumulation. Our model explains many features of rising wage inequality in the U.S. economy (Heckman et al., 1998). In this paper, we use our model to study the impacts on skill formation of proposals to switch from progressive taxes to flat income and consumption taxes. For the sake of brevity, we focus on steady states in this paper, although we study both transitions and steady states in our research.

I. Our Model

Our analysis builds on the model of Alan Auerbach and Laurence Kotlikoff (1987) in two ways: (i) we introduce skill formation and consider both schooling choices and investment in on-the-job training; and (ii) we allow for heterogeneity in ability, endowments and skills. Different schooling levels are associated with different skills and different post-school investment functions.

We relax their efficiency-units assumption for labor services. Models with efficiency units for labor services do not explain rising wage inequality among skill groups. Our model has three sources of heterogeneity among persons: (i) in age; (ii) in ability to learn and in initial endowments; and (iii) in the economic histories experienced by cohorts. In a transition period, different cohorts face different skill prices, make different investment decisions, and hence, accumulate different amounts of human capital and have different wage levels and trajectories. Our model extends the analysis of James Davies and John Whalley (1991) who introduce human capital into the Auerbach-Kotlikoff model but assume only one skill. We allow for multiple skills, incorporate both schooling and on-the-job training, and allow for rational expectations in calculating transition paths.

In our model, individuals live for \bar{a} years and retire after $a_R \leq \bar{a}$ years. In the first stage of the life cycle, a prospective student chooses the schooling option that gives him the highest level of lifetime utility. Define K_{at} as the stock of physical capital held at time t by a person age a ; H_{at}^S is the stock of human capital at time t of type S at age a . The optimal-life-cycle problem can be solved in two stages. First, condition on schooling and solve for the optimal path of consumption (C_{at}) and post-school investment time (I_{at}^S) for each schooling level. Second, let individuals select among schooling levels to maximize lifetime welfare.

Given S , an individual age a at time t has the following value function:

$$\begin{aligned} (1) \quad & V_{at}(H_{at}^S, K_{at}, S) \\ &= \max_{C_{at}, I_{at}^S} \frac{C_{at}^\gamma - 1}{\gamma} \\ &+ \delta V_{a+1,t+1}(H_{a+1,t+1}^S, K_{a+1,t+1}, S) \end{aligned}$$

* Heckman and Lochner: Department of Economics, University of Chicago, 1126 E. 59th Street, Chicago, IL 60637; Taber: Department of Economics, Northwestern University, 2003 Sheridan Road, Evanston, IL 60208. We thank Ken Judd and Jim Poterba for helpful comments. This research was supported by grants from the Russell Sage Foundation and NSF grant SBR-93-21-048.

where δ is a time preference discount factor. We follow Kotlikoff et al. (1997) by assuming that the tax schedule can be approximated by a progressive tax on labor income and a flat tax on capital income. This gives the following dynamic budget constraint:

$$(2) \quad K_{a+1,t+1} \leq K_{a,t}[1 + (1 - \tau_k)r_t] + R_t^S H_{a,t}^S (1 - I_{a,t}^S) - \tau_\ell [R_t^S H_{a,t}^S (1 - I_{a,t}^S)] - C_{a,t}$$

where τ_k is the proportional tax rate on capital, τ_ℓ is the progressive tax schedule on labor earnings, R_t^S is the price of human capital services of type S at time t , and r_t is the net return on physical capital at time t . We experiment with other progressive tax schedules and obtain results similar to the ones we report here. In this paper, we abstract from labor supply. Estimates of intertemporal substitution in labor supply estimated on annual data are small, so ignoring labor supply does not affect our analysis. This simplification makes our model comparable to that of Davies and Whalley (1991), who also ignore leisure.

On-the-job human capital for a person of schooling level S accumulates through the human-capital production function:

$$(3) \quad H_{a+1,t+1}^S = A^S(\theta)(I_{a,t}^S)^{\alpha_S}(H_{a,t}^S)^{\beta_S} + (1 - \sigma^S)H_{a,t}^S$$

where the conditions $0 < \alpha_S < 1$ and $0 \leq \beta_S \leq 1$ guarantee that the problem is concave, and σ^S is the rate of depreciation of skill (S)-specific human capital. This functional form is widely used in both the empirical literature and the literature on human-capital accumulation. The parameters α and β are also permitted to be S -specific, which emphasizes that schooling affects the process of learning on the job in a variety of different ways.

Notably absent from our model are short-run credit constraints which are often featured

in the literature on schooling and human capital accumulation. Our model is consistent with the evidence presented in Stephen Cameron and Heckman (1998) that long-run family factors correlated with income [the θ operating through $A^S(\theta)$ and the initial condition for (3)] affect schooling but that short-term credit constraints are not empirically important. Such long-run factors account for the empirically well-known correlation between schooling attainment and family income.

At the beginning of life, agents choose the value of S that maximizes lifetime utility:

$$(4) \quad \hat{S} = \underset{S}{\operatorname{argmax}} [V^S(\theta) - D^S + \varepsilon^S]$$

where $V^S(\theta)$ is the tax-adjusted present value of earnings at schooling level S computed from the optimal program, D^S is the discounted tuition cost of schooling, and ε^S represents nonpecuniary benefits expressed in present-value terms.

Tuition costs are permitted to change over time so that different cohorts face different schooling costs. The economy is assumed to be competitive so that the prices of skills and capital services are determined as derivatives of an aggregate production function. In order to compute service-flow prices for capital and the different types of human capital, it is necessary to construct aggregates for each of the factors over each of the ability types and over all cohorts to insert into an aggregate production function.

Human capital of type S is a perfect substitute for any other human capital of the same schooling type, whatever the age or experience level of the agent, but it is not perfectly substitutable with human capital from other schooling levels. In our model, cohorts differ from each other because they face different price paths and policy environments within their lifetimes.

Our aggregate production function exhibits constant returns to scale. The equilibrium conditions require that marginal products equal pretax prices. In the two-skill economy we analyze, the production function at time t is defined over the inputs \bar{H}_t^1 , \bar{H}_t^2 , and \bar{K}_t , where \bar{H}_t^1 and \bar{H}_t^2 are aggregates of *utilized* skills

(high school and college, respectively) supplied to production, and \bar{K}_t is the aggregate stock of capital. The technology we use is

$$F(\bar{H}_t^1, \bar{H}_t^2, \bar{K}_t) \\ = a_3 \{ a_2 [a_1 (\bar{H}_t^1)^{\rho_1} + (1 - a_1) (\bar{H}_t^2)^{\rho_1}]^{\rho_2/\rho_1} \\ + (1 - a_2) \bar{K}_t^{\rho_2} \}^{1/\rho_2}.$$

We estimate that $\rho_2 = 0$ but $\rho_1 = 0.693$, which yields an elasticity of substitution between high school and college human capital of 1.441.

Human-capital accumulation functions (3) are estimated using micro data assuming that taxes are proportional. However, an extensive sensitivity analysis reveals that, within the range of the data for the U.S. economy, misspecification of the tax system does not affect parameter estimates if the model is recalibrated on aggregate data. We now use the model to investigate tax policies.

II. Tax Effects on Human-Capital Accumulation

In the absence of labor-supply and direct pecuniary or nonpecuniary costs of human-capital investment, there is no effect of a proportional wage tax on human-capital accumulation. Both marginal returns and costs are scaled down in the same proportion. When untaxed costs or returns to college are added to the model (i.e., nonpecuniary costs/benefits), proportional taxation is no longer neutral. An increase in the tax rate decreases college attendance if the net financial benefit before taxes is positive ($V^2 - D^2 - V^1 > 0$). Progressivity reinforces this effect. A progressive wage tax reduces the incentive to accumulate skills, since human capital promotes earnings growth and moves persons to higher tax brackets. As a result, marginal returns on future earnings are reduced more than marginal costs of schooling.

Heckman (1976) notes that in a partial-equilibrium model, proportional taxation of interest income with full deductibility of all borrowing costs reduces the after-tax interest rate and, hence, promotes human-capital accumulation. In a time-separable, representative-agent general-equilibrium model,

the after-tax interest rate is unaffected by the tax policy in steady state as agents shift to human capital from physical capital (see Philip Trostel, 1993). In that framework, flat taxes with full deductibility have no effect on human-capital investment. In a dynamic overlapping-generations model with heterogeneous agents and endogenous skill formation and with progressive rates, taxes have ambiguous effects on human capital, and both their quantitative and qualitative effects can only be resolved by empirical research. We use our empirically grounded model to study alternative proposals for tax reform.

III. Analyzing Two Tax Reforms

Following Kotlikoff et al. (1997), we assume that the U.S. income tax can be captured by a progressive tax on labor income and a flat tax on capital income. Each earner has 1.22 children and is single. For each additional dollar beyond \$9,660, there is an increase in itemized deductions of 7.55 cents. An individual with labor income Y has taxable income $(Y - 9,660)(1 - 0.0755)$. Using the 1995 tax schedule, we compute the taxes paid by income and approximate this schedule by a second-order polynomial. We assume a 0.15 flat tax rate on physical capital.

We consider two revenue-neutral tax reforms from this benchmark progressive schedule. The first reform (which we call "flat tax") is a revenue-neutral flattening of the tax on labor earnings, holding the initial flat tax on capital income constant. The second reform ("flat consumption tax") is a uniform flat tax on consumption. In both flat-tax schemes, tuition is not treated as deductible. (We discuss the consequences of making it deductible below.) For each tax, we consider two models: (i) a partial-equilibrium model in which skill prices and interest rates are fixed and (ii) a closed-economy general-equilibrium model where skill prices and interest rates adjust.

Table 1 presents both partial-equilibrium and general-equilibrium results measured relative to a benchmark economy with the Kotlikoff et al. (1997) tax schedule. We first discuss the partial-equilibrium effects of a move to a flat tax, which eliminates progressivity in wages and stimulates skill formation.

TABLE 1—COMPARISON OF STEADY STATES
UNDER ALTERNATIVE TAX REGIMES

Variable	Percentage difference from benchmark progressive case			
	Flat tax		Flat consumption tax	
	PE	GE	PE	GE
After-tax interest rate	0.00	1.96	17.65	3.31
Interest rate	0.00	1.96	0.00	-12.18
Skill price, college human capital	0.00	-1.31	0.00	3.38
Skill price, high-school human capital	0.00	-0.01	0.00	4.65
Stock of physical capital	-15.07	-0.79	86.50	19.55
Stock of college human capital	22.41	2.82	-15.77	1.85
Stock of high-school human capital	-9.94	0.90	1.88	0.08
Stock of college human capital per college graduate	3.04	2.55	-4.08	1.72
Stock of high-school human capital per high-school graduate	1.84	1.07	-5.23	0.16
Fraction attending college	18.79	0.26	-12.18	0.13
Aggregate output	-0.09	1.15	15.76	4.98
Aggregate consumption	-0.08	0.16	7.60	3.66
Mean wage, college graduate	3.39	2.60	0.12	6.96
Mean wage, high school graduate	2.44	2.44	0.25	6.82
Standard deviation, log wage	4.09	1.56	-1.94	0.69
College/high-school wage premium ^a	1.92	-0.45	3.10	0.18

Notes: In the progressive case, we allow for a progressive tax on labor earnings but assume a 15-percent flat tax on capital. In the flat-tax regime, we hold the tax on capital fixed at 15 percent but assume that the tax on labor income is flat. Balancing the budget yields a tax rate on labor income of 7.7 percent. In the consumption-tax reform, only consumption is taxed at a 10-percent rate. PE = partial equilibrium; GE = general equilibrium.

^a The college/high-school wage premium measures the difference in mean log wage rates between college graduates and high-school graduates with ten years of work experience.

College attendance rises dramatically as the higher earnings associated with college graduation are no longer taxed away at higher rates. The amount of post-school on-the-job training also increases for each skill group (as measured by the stocks of human capital per worker of each skill). The aggregate stock of high-school human capital declines, while the aggregate stock of college human capital increases as a result of the rise in college enrollment. The college-high-school wage differential increases slightly as does another widely used measure of inequality, the standard deviation of log wages. The effects of reform on aggregates of consumption and output are modest at best. However, capital formation is greatly reduced as the tax code now favors human capital compared to the benchmark economy.

In general equilibrium, the effects of the reform on skill formation are, in general, qualitatively similar, but they are greatly diminished. The effects on aggregate consumption and output are weak, as they are in the partial-equilibrium case. Furthermore, the negative effects of the reform on physical capital are muted, since the return to capital increases. The rise in the after-tax interest rate chokes off skill investment. Per capita post-school on-the-job training accumulation still increases for both skill groups, although the increase is dampened compared to the partial-equilibrium case. Aggregate stocks of both high-school and college human capital now rise, since college enrollment increases much less. The distinction between partial equilibrium and general equilibrium is especially striking for the fraction attending college. Though not shown in the table, college attendance increases only for the most able, whereas in the partial-equilibrium case, it increases for all ability groups. Changes in skill prices and interest rates virtually offset the removal of the disincentives of progressive taxes on schooling enrollment. The college-high-school wage differential (at 10 years of experience) now declines slightly, and the increase in the standard deviation of log wages is less. In general equilibrium, the increase in the standard deviation is smaller, because skill prices adjust and because higher after-tax interest rates flatten wage profiles.

Next, consider a move to a flat consumption tax. This reform is more pro-capital and is less favorable to human capital. It raises output, capital, and consumption more than a flat-tax reform, and it reduces the aggregate stock of high-skill human capital and the stock of human capital per worker for each skill group. The fraction attending college declines. The reform raises wage inequality as measured by the college-high-school wage premium but lowers it as measured by the standard deviation of log wages.

In general equilibrium, this reform is slightly less favorable to human-capital formation than the flat tax, since the after-tax rate of return on capital rises more. College attendance increases slightly, but the increase is concentrated among the least and most able persons. Wage inequality increases slightly by both conventional measures. Real wages rise for both skill groups. The effect is greater than in the flat-tax reform. This is due to a larger increase in capital under proportional consumption taxation. Since capital is a direct complement with both forms of human capital, the increase in capital raises skill prices about equally for both skill groups. The greater increase in real wages in this case is not due to a larger increase in per capita human-capital accumulation within skill groups.

When we introduce deductibility of tuition in both reforms and preserve revenue-neutrality, there is virtually no effect on skill formation (or anything else) in general equilibrium. This is consistent with our other work in which we show that general-equilibrium effects of tuition subsidies are small. The lessons from partial-equilibrium analyses are substantially misleading guides in analyzing the effects of tax and tuition policy on skill formation. Changes to proportional taxation are unlikely to have large effects on skill formation or output. A change to a flat consumption tax has the largest effect on output,

consumption, and real wages, but it also slightly raises wage inequality. These conclusions also hold for open-economy simulations in which the interest rate is set in world markets. They are robust to a variety of tax schedules and empirically grounded parameter estimates.

REFERENCES

- Aaron, Henry and Gale, William. *Economic effects of fundamental tax reform*. Washington, DC: Brookings Institution Press, 1996.
- Auerbach, Alan and Kotlikoff, Laurence. *Dynamic fiscal policy*. Cambridge: Cambridge University Press, 1987.
- Cameron, Stephen and Heckman, James. "Life Cycle Schooling and Educational Selectivity: Models and Evidence." *Journal of Political Economy*, April 1998, 108(2), pp. 262-334.
- Davies, James and Whalley, John. "Taxes and Capital Formation: How Important Is Human Capital?" in B. Bernheim, and J. Shoven, eds., *National saving and economic performance*. Chicago: University of Chicago Press, 1991.
- Heckman, James. "A Life Cycle Model of Earnings, Learning and Consumption." *Journal of Political Economy*, August 1976, 84(4), part 2, pp. S11-S44.
- Heckman, James; Lochner, Lance and Taber, Christopher. "Explaining Rising Wage Inequality: Explorations with a Dynamic General Equilibrium Model of Labor Earnings with Heterogeneous Agents." *Review of Economic Dynamics*, 1998 (forthcoming).
- Kotlikoff, Laurence; Smetters, Kent and Walliser, Jan. "The Economic Impact of Privatizing Social Security." Unpublished manuscript, Boston University, 1997.
- Trostel, Philip. "The Effect of Taxation on Human Capital." *Journal of Political Economy*, April 1993, 101(2), pp. 327-50.

Does Government R&D Policy Mainly Benefit Scientists and Engineers?

By AUSTAN GOOLSBEE*

Substantial evidence has shown that the social rate of return to R&D spending significantly exceeds the private rate of return (see Zvi Griliches, 1991; Bronwyn Hall, 1996), and conventional wisdom holds that this public-good nature of inventive activity makes private R&D spending lower than the social optimum and warrants a role for government involvement to increase it. In the United States, government funds have provided a large fraction, often more than half, of the entire nation's R&D spending.

This paper, however, suggests a practical difficulty with government efforts to increase inventive activity. Specifically, most R&D spending is actually salary payments for R&D workers, and the supply of this scientific and engineering talent is quite inelastic. When the government increases R&D spending through subsidies or by direct provision, a significant fraction of the increased spending goes directly into higher wages, an increase in the price rather than the quantity of inventive activity. The conventional literature, by looking at total R&D spending, misses this distinction. The magnitudes found in this paper suggest that the conventional literature may overstate the effects of government R&D spending by as much as 30–50 percent. In this sense, R&D policy may be less about increasing innovation and more about rewarding the human capital of scientists. The results also imply that, by altering the wages of scientists and engineers even for firms not receiving federal support, government funding crowds out private inventive activity.

The asset-market effects of government R&D policy clearly parallel other asset-market incidence results such as those explored by Goolsbee (1997, 1998) showing that capital-investment subsidies are capitalized into the prices of equipment and the wages of capital-goods workers or James Poterba (1984) showing that housing subsidies raise house prices.

I. The Government and R&D Spending

Since World War II, the government has been an extremely important part of R&D spending in the United States, and a large academic literature has evaluated the effectiveness of many government R&D policies (see Hall, 1996). The government has indirectly supported R&D with the patent system, the research and experimentation (R&E) tax credit, and credits for R&D through multinational tax rules and has also directly provided R&D spending through universities, NASA, the Department of Defense, the Department of Energy, the Department of Health, and other specific programs (see the data in NSF's *Research and Development in Industry* [National Science Foundation, 1995] for an overview).

These R&D programs are costly. R&D spending has comprised between 2 percent and 3 percent of GDP since the 1960's, and the federal government's share has consistently been between one-third and two-thirds of the total. In 1995, it amounted to almost \$70 billion in direct funding. Most of the federal money has gone to the Defense Department (always more than half, and up to 70 percent in some years) and since the 1960's, the biggest variations in government R&D spending have been fluctuations in the defense component. Such R&D spending was high in the late 1960's, fell through the 1970's, rose in the 1980's, and fell again in the 1990's. This variation can be used to identify the effect of R&D spending on the market for scientists and engineers.

* University of Chicago, Graduate School of Business, 1101 E. 58th Street, Chicago, IL 60637, American Bar Foundation, and National Bureau of Economic Research. I thank Pete Klenow, Jim Poterba, Harvey Rosen, and Nancy Stokey for helpful comments. This research is funded by the University of Chicago GSB and by a grant from the American Bar Foundation.

II. The Market for Scientists and Engineers

Although the NSF data on R&D spending are widely used in the literature, minimal attention has been paid to where the data show the money goes. Some work has examined the composition of R&D spending going to basic versus applied research, but little has focused on the share going to scientists. The National Science Foundation (1995) documents that between 45 percent and 83 percent of total spending is wages and benefits of scientific personnel (depending on how one counts overhead, which includes individual benefits). A reasonable approximation for the total share might be two-thirds. Thus, when the data show that in 1995 the government spent almost \$70 billion on R&D, \$45 billion of that was wages and benefits for R&D workers.

The supply of R&D workers, however, is inelastic. Scientists and engineers have extremely high human capital that takes many years to accumulate, and entry is small. The data in Jaewoo Ryoo and Sherwin Rosen (1992) show that the biggest graduating classes generated only about 4,000 engineering Ph.D.'s, 20,000 M.S.'s, and about 75,000 B.S.'s (a modest amount relative to the 1.4 million stock of engineers), and it is likely that the supply of ideas is probably even less elastic than the supply of personnel.

If government R&D spending goes predominantly to scientific labor and the supply of that labor is inelastic, it means that the inventive value of a dollar of R&D spending will vary over time depending on demand conditions and that government spending will translate into higher wages.

III. The Impact of R&D Funding on R&D Workers

To show that this is an important issue in practice, I turn to the data in the Current Population Survey on the income of scientists and engineers from 1968 to 1994. I restrict the sample to include only full-time (35 or more hours per week), white, male engineers and scientists who are between the ages of 21 and 65 and have at least a college degree. These restrictions eliminate little of the sample because this has been a very white, male,

TABLE 1—THE EFFECT OF R&D ON INCOME AND HOURS OF ENGINEERS AND SCIENTISTS

Dependent variable	$\frac{R\&D}{GDP}$	$\frac{R\&D_{fed}}{GDP}$	Mean	R^2
ln(income)	0.300 (0.038)			0.288
ln(income)		0.232 (0.030)		0.288
ln(wage)		0.219 (0.032)		0.255
ln(hours)		0.009 (0.013)		0.012
ln(income)		0.094 (0.043)	0.171 (0.038)	0.277

Notes: Each row presents the R&D coefficient from a regression including experience, the square of experience, dummies for marriage, scientific occupation, post-college education, GDP growth, and a time trend. The dependent variable is listed in the first column. The sample in each case is 1968–1994. There are 17,700 observations in each regression, and the standard errors are in parentheses. The second column is the ratio of R&D to GDP. The second column is the ratio of federal R&D to GDP. The mean variable is the ratio of the log of the mean federal R&D to GDP for the four previous years.

college-degree-dominated field. The sample includes 17,700 individuals.

The first row of Table 1 presents a regression for the log of real income on the log of total R&D spending as a share of GDP, the GDP growth rate to represent business-cycle fluctuations, a dummy equal to 1 if a person has greater than a college degree, a dummy equal to 1 if the individual is married, a dummy equal to 1 if the individual is in a scientific (nonengineering) occupation, experience, experience-squared (experience is defined as age minus 22 for people with a college degree and as age minus 26 for people with an advanced degree), and a time trend. For reasons of space, I report only the R&D coefficients (the other coefficients had conventional magnitudes and signs).

The R&D coefficient clearly shows that higher spending increases the incomes of scientists and engineers. A one-standard-deviation increase in R&D spending (10 percent) would increase incomes by about 3 percent. The National Science Foundation (1995) estimates that there were 768,500 full-time-equivalent R&D workers in 1994, but

Occupational Employment Statistics reports 1,722,000 who describe their occupation as physical scientist, life scientist, or engineer. The results here imply that wages rise for all scientists and engineers, but I cannot distinguish between wages rising by 3 percent for everyone and wages rising by 6.75 percent for the 45 percent actually engaged in R&D and zero for the others.

Since the issue of the paper specifically concerns the effectiveness of government policy, the second row uses the log of the ratio of federally funded R&D to GDP. The coefficient is slightly smaller but still highly significant. Other specifications using income relative to average manufacturing earnings or the real level of R&D spending (rather than as a fraction of GDP) gave similar results.

The third and fourth rows break the effect into a wage and an hours component. The wage variable used in the third row is defined as annual income divided by 50 times the number of hours last week (i.e., assuming two weeks vacation) and shows that wages account for about 95 percent of the income increase. The hours variable used in the fourth row is the log of hours worked last week, and the coefficient is extremely small and is not significant (accounting for censoring with a Tobit gave identical results).

Together these results strongly suggest that R&D spending increases wages and not effort. Directly estimating a short-run supply curve by regressing the log of hours on the log of wages and demographic factors while instrumenting for wages with R&D spending gave elasticities of supply between 0.1 and 0.2, which were highly significant.

The last row looks at the longer-run effect of R&D spending on salaries by including not just current federal R&D spending, but also the average level of the previous four years. There is no reduction in the salary gains over this longer period. In fact, a permanent one-standard-deviation increase in R&D spending raises wages 1 percent immediately and an additional 2 percent over the next four years. Thus the supply of scientific personnel does not seem to be especially elastic in the medium-run either.

Because the measures of R&D are annual variables, it is not possible to include year

TABLE 2—RESULTS DISAGGREGATED BY OCCUPATION

Occupation	Regression coefficient	<i>n</i>	<i>R</i> ²
Engineer			
Aeronautical	0.447 (0.122)	723	0.354
Mechanical	0.380 (0.088)	2,186	0.276
Metallurgical	0.375 (0.220)	205	0.312
Electrical	0.180 (0.062)	3,693	0.305
Chemical	0.077 (0.155)	706	0.254
Industrial	0.061 (0.092)	1,454	0.265
Civil	0.021 (0.086)	1,996	0.264
Mining	-0.148 (0.353)	103	0.295
Scientist			
Physical	0.564 (0.202)	328	0.461
Geological	0.192 (0.226)	436	0.294
Biological	0.169 (0.204)	497	0.312
Agricultural	0.157 (0.281)	194	0.384

Notes: Each row presents the results of the regression in row 2 of Table 1 but for a single occupation. The total does not sum to 17,700 because occupations with classification changes are not included. The dependent variable in each is the log of real income. The sample in each case is 1968–1994. Standard errors are in parentheses.

dummies to control for other important time-series variables which are unobserved and possibly correlated with federal spending. The results in Table 2, however, suggest that omitted variables are not causing the estimated response. The table breaks the scientists and engineers into groups by specific occupation for those occupations that remain consistent over the full sample including aeronautical, chemical, civil, electrical, industrial, metallurgical, and mining engineers, agricultural scientists, biologists, geologists, and physicists. Since federal R&D spending has had a heavy defense and space focus (85–95 percent of federal spending went to defense, space, and energy throughout the sample), if R&D spending is the true source of salary increases, salaries for aeronautical engineers and physicists might rise more than for civil engineers and agricultural scientists.

The results clearly show this. Federal R&D spending has no significant effect on the salaries of mining, civil, industrial, and chemical engineers (occupations in areas with little federal R&D spending) but has a major impact on the salaries of aeronautical, mechanical, metallurgical, and electrical engineers. Among scientists, physicists are the major beneficiar-

ies. A further regression (not reported in the table) found that, predictably, biologists benefit most from the health component of federal R&D.

IV. Implications

The first implication of the results is that evaluations of government R&D policy may significantly overstate their effects on inventive activity. The defense buildup during Ronald Reagan's administration, for example, increased federal R&D spending as a share of GDP by 11 percent from 1980 to 1984. The results here imply that this would increase the salaries of scientists and engineers by 3.3 percent and, at a more disaggregated level, increase the salaries of physicists by 6.2 percent, aeronautical engineers by almost 5 percent, mechanical and metallurgical engineers by more than 4 percent, and electrical engineers more than 2 percent.

Taking the wage increase at 3.3 percent and the wage share of R&D spending at two-thirds, the effect on the true "quantity" of R&D is not the 11-percent increase in total R&D spending but rather 7.6–30-percent smaller. If all the spending went to physicists, aeronautical engineers, and mechanical engineers, the true quantity increase would be even smaller: as much as 40–50-percent lower than the spending increase would indicate. This same reasoning could similarly reduce estimated impacts of the R&E tax credit on true inventive activity such as that of Hall (1993).

The second implication is that government R&D directly crowds out private inventive activity. By increasing the wages of electrical engineers, for example, government R&D spending increases R&D costs for computer manufacturers who do not receive federal funding but are forced to pay their scientists higher wages. The simple correlation between the ratio of federally funded R&D to GDP and the ratio of nonfederally funded R&D to GDP is -0.4 , so this may be important.

This general-equilibrium type of effect may also help explain why the response of R&D spending to R&D tax subsidies is usually estimated to be larger using cross-sectional data than using aggregate data (see the evidence discussed in Hall [1993]). If subsidies raise

the wages of R&D workers and thereby shift R&D spending away from firms that cannot use the subsidy toward the firms that can, this will make the cross-sectional elasticity large even if the aggregate elasticity is small. This response is consistent with the evidence of Philip Berger (1993) that R&D spending among firms that cannot use R&D tax subsidies falls when the subsidies rise.

V. Conclusion

The evidence shows that a major component of government R&D spending is windfall gains to R&D workers. Incomes rise significantly while hours rise little, and the increases are concentrated within the engineering and science professions in exactly the specialties most heavily involved in federal research. The implication for the evaluation of government R&D policy is that government spending is likely to have significantly smaller effects on the quantity of inventive activity than is implied by looking at R&D spending alone, and also that government spending directly crowds out private spending by raising wages.

The evidence means that federal R&D policy is, in large measure, just a way of subsidizing scientific human capital and its acquisition. The importance of increasing the share of the labor force engaged in R&D should not be downplayed, however. Endogenous growth theory (e.g., Paul Romer, 1990) has focused attention on the potential importance of R&D for growth, and some evidence (Kevin Murphy et al., 1991) indicates that encouraging people to become engineers may help society simply by reducing the number of lawyers! Ryoo and Rosen (1992) have shown that college engineers are quite elastic in their choices of majors.

One should recognize, however, that there may be better ways to encourage scientific careers than by subsidizing the wages of all scientific personnel through R&D spending, just as it can cost less to subsidize corporate investment than to reduce the corporate income tax. Whatever the motivation, the results in this paper make clear that the difference between R&D spending and true inventive activity should be factored into any analysis of government R&D policy.

REFERENCES

- Berger, Philip. "Explicit and Implicit Tax Effects of the R&D Tax Credit." *Journal of Accounting Research*, Autumn 1993, 31(2), pp. 131-71.
- Goolsbee, Austan. "To the Workers Go the Spoils? The Incidence of Investment Tax Subsidies." American Bar Foundation (Chicago, IL) Working Paper No. 9611, 1997.
- . "Investment Tax Incentives, Prices, and the Supply of Capital Goods." *Quarterly Journal of Economics*, February 1998, 113(1), pp. 121-48.
- Griliches, Zvi. "The Search for R&D Spillovers." National Bureau of Economic Research (Cambridge, MA) Working Paper No. 3768, 1991.
- Hall, Bronwyn. "R&D Tax Policy During the 1980s: Success or Failure?" James Poterba, ed., *Policy and the economy*, Vol. 7. Cambridge, MA: MIT Press, 1993, pp. 1-35.
- . "The Private and Social Returns to Research and Development," Bruce Smith and Claude Barfield, eds., *Technology, R&D, and the economy*. Washington, DC: Brookings Institution Press, 1996, pp. 1-14.
- Murphy, Kevin; Shleifer, Andrei and Vishny, Robert. "The Allocation of Talent: Implications for Growth." *Quarterly Journal of Economics*, May 1991, 106(2), pp. 503-30.
- National Science Foundation. *Research and development in industry*. Washington, DC: U.S. Government Printing Office, 1995.
- Poterba, James M. "Tax Subsidies to Owner-Occupied Housing: An Asset Market Approach." *Quarterly Journal of Economics*, November 1984, 99(4), pp. 729-52.
- Romer, Paul. "Endogenous Technological Change." *Journal of Political Economy*, October 1990, 98(5), Part 2, pp. S71-S102.
- Ryoo, J. and Rosen, S. "The Market for Engineers." Stigler Center Paper No. 83, University of Chicago, 1992.

RETHINKING PUBLIC EDUCATION[†]

The Origins of State-Level Differences in the Public Provision of Higher Education: 1890–1940

By CLAUDIA GOLDIN AND LAWRENCE F. KATZ *

In recent years, the publicly controlled sector has accounted for about 67 percent of all students in four-year institutions of higher education, whereas in 1897 the figure was only 22 percent.¹ The transition was early and swift, however, for on the eve of the United States' entry into World War II almost 50 percent of all students were in public-sector institutions. Thus, by 1940 more than 60 percent of the century-long expansion in public-sector enrollments (four-year) had already taken place. The relative rise of public higher education occurred decades before the rapid postwar acceleration of college enrollments and is apparent even without including community colleges.

Differences across states in public-sector enrollments and in public support for higher education were considerable in the past but narrowed markedly after 1950. In 1929, enrollments in publicly controlled institutions averaged 6 per 1,000 residents in the Mountain

and Pacific states, but only 0.8 per 1,000 residents in New England. State and local spending on higher education was \$2,057 per 1,000 inhabitants in the Mountain states, but just \$458 per 1,000 inhabitants in New England. In 1900 the cross-state coefficient of variation of public-sector enrollments per capita was 1.14, but this fell to 0.60 by 1929 and stands today at about 0.21. The convergence for the private-sector figure was far less (0.88 in 1900 and 0.70 in 1994). Changes in the cross-state variation in spending per capita and in public-sector tuition are similar.

Despite the potent forces of market integration (Caroline Hoxby, 1997), many higher-education indicators exhibit substantial long-term persistence. A strong relationship exists at the state level between private-sector enrollments per capita today and those in 1900 ($\rho = 0.64$). A strong correlation ($\rho = 0.46$) also exists between public higher-education spending per capita today and the year of statehood (see also John M. Quigley and Daniel L. Rubinfeld, 1993). States admitted earlier have stronger private sectors and less well-supported public sectors, and the relationship is more complex than a simple difference between the Northeast and the Far West.

Both the public and private higher-education sectors evolved in the 1890–1940 period. Fundamental changes in the creation and diffusion of knowledge, such as the specialization of disciplines, the professionalization of many occupations, the secularization of higher education, the increased role of research, and the ascent of the “university” form, altered the industrial structure of higher education. These changes increased the scale and scope of institutions and gave a productivity advantage to certain public-sector institutions.

[†] *Discussants:* Robert P. Inman, University of Pennsylvania; Christopher S. Jencks, Harvard University.

* Department of Economics, Harvard University, Cambridge, MA 02138. We acknowledge generous support from the National Science Foundation for our research and from the Russell Sage Foundation for our 1997–1998 academic leave. We thank Cheryl Seleski and Kerry Woodward for their superb assistance in the coding of the higher-education data set. Conversations with Caroline Minter Hoxby, James Poterba, and Marcus Stanley have helped clarify some of the ideas in this paper. We thank them all.

¹ In the interest of brevity, all citations for recent data are U.S. Department of Education (1996) and “today” means 1994. For all other data sources, see Goldin and Katz (1998). The share of total enrollments accounted for by publicly controlled institutions in 1994 was 78 percent when two-year institutions are also included. The term “public sector” in this paper means “publicly controlled sector.”

Thus history matters in understanding the higher-education industry. We look here at the origins of state-level differences in support for public-sector higher education. We address why higher-education enrollments increased more in the public than in the private sector long before the great advance in college enrollments and why certain states historically have provided more generous support to their public institutions of higher education.

I. The Evolution of Higher Education to 1940: Supply and Demand Factors

The relative growth of enrollments in public-sector institutions was but one part of a larger set of changes that swept knowledge creation and diffusion around the turn of the 20th century (see Laurence R. Veysey, 1965). During the early to mid-19th century, institutions of higher education were often staffed by a handful of faculty, each of whom taught several subjects. The diffusion of the scientific method, practically oriented courses, the "lecture" method of teaching, and disciplinary specialization fundamentally altered the knowledge industry (see e.g., Alexandra Oleson and John Voss, 1979). An intricate division of labor began to permeate higher education, and with it came greater economies of scale in the production of higher-education services. But more important to the story at hand is that the diffusion of knowledge became closely bound up with the *creation* of knowledge. Research became the handmaiden of teaching that we believe it is today.

Three additional changes, somewhat separate in their causes, added to the altered structure of the higher-education industry and to the relative growth of the public sector. One was the secularization of the college. A second was the growth of professional schools. The last, and most important, was the emergence of an overarching, umbrella organization known as the "university" and the integration of professional schools into it.

A "university" is not just a department store of education services brought together for the benefit of its student-clients. Rather, the modern university is a production center in which the research of one part enhances

the teaching and research of the other parts. The "university" form was an organizational innovation enabling the exploitation of technical complementarities among its various components. The public sector had a disproportionate share of universities early in the period under study, and they grew faster than did those in the private sector before World War II. The fact that the publicly controlled sector was disproportionately established in the university, research-oriented form gave it a substantial advantage before 1940.

Certain universities had, as well, the capacity to bestow reputation on new divisions in untried areas, such as business schools, and in areas plagued by claims of quackery, as were medical schools in the wake of the 1910 Flexner Report. Thus, the university came to have all three types of characteristics: those of the "department store," the "integrated factory," and the "brand name."

An essential ingredient in the changing nature of higher education was the enormous increase in the supply of potential clients after 1910 with the rise in graduation rates in American high schools. In 1910, less than 10 percent of young Americans graduated from public and private secondary schools, but by the mid-1930's about 50 percent did in most states outside the South (Goldin, 1998). Moreover, the increase was greater in states having well-functioning public higher-education institutions. High-school graduation rates expanded more rapidly from 1910 to 1928 in states with larger public-sector undergraduate enrollment rates in 1910 (Goldin and Katz, 1997).

II. Changes in the Structure of the Higher-Education Sector

Educational institutions, in both the public and the private sectors, expanded in scale and scope in the half century before 1940.² In 1897 the median private institution had 128 students, and the median private student was in

² We exclude, in this discussion, all independent teachers' colleges, two-year colleges, and students in the preparatory departments of higher-educational institutions.

an institution with 505 students. In the public sector these two measures were 242 and 787. Publicly controlled institutions in 1897 were, on average, larger than private ones, but they were not very much larger, and the largest were private. By 1923 the average institution in both sectors had grown substantially, but public-sector institutions had grown far more. The median private institution had 357 students, and the median student in the private sector was in an institution with 1,685 students; the public-sector numbers were 1,225 and 3,950. The average public-sector institution grew by a factor that was 1.8 times the corresponding factor for the private-sector. By the 1920's, public-sector institutions of higher education had become large, research-oriented universities.

Anecdotal evidence, bolstered by quantitative results given below, suggests that, in states having a concentration of economic activity by industrial-output, farm-product, mining, or oil interests, the public sector invested heavily in training and research. Wisconsin did research on dairy products, Iowa on corn, Colorado and other Western states on mining, North Carolina on tobacco, and Oklahoma and Texas on oil exploration and refining (Nathan Rosenberg and Richard R. Nelson, 1993). States subsidized the training of professionals for whom they perceived a need. In 1923 engineers were 15 percent of all students in four-year public institutions but only 5.4 percent in the private sector, and 60 percent of all engineers were trained in the state sector. State institutions of higher education attained the status of "university" to a greater extent than did privately controlled institutions, for they had all the component parts *and* state research funds when funds were less available elsewhere.

III. What Accounts for the Relative Growth of Public Higher-Education Institutions?

From 1897 to 1940 the fraction of students in publicly controlled institutions increased from 0.22 to just below 0.5 including junior colleges, and to around 0.45 excluding them. From 1933 to 1990 the series, which includes teachers' colleges but excludes junior colleges, increased from around 0.5, just before

World War II, to 0.67 in 1990.³ The full century, then, experienced an relative increase in public-sector enrollment from 22 percent to 67 percent of total enrollment.

What accounted for the relative growth in public-sector enrollment? Our evidence is consistent with two complementary interpretations. One emphasizes the changed structure of the higher-education industry, whereas the other features the increased supply of potential college students stemming from the high-school movement.

The more novel of the explanations we offer is that the application of the scientific method and the increased division of labor and specialization in higher education disproportionately benefited certain types of institutions. Those that had access to research funds were initially large and diverse, were nonsectarian, and had both reputation and a long-purse were in the best position to prosper from the changes. Most public-sector institutions, and some in the private sector, were so situated and thus flourished and expanded relative to others in the wake of various "technological" changes that shook higher education from around 1880 to 1910.

Consider one example of the phenomenon we believe to have operated (see also Goldin and Katz [1998]). In the late 19th century a large fraction of individuals in engineering, law, business, medicine, agricultural science, and chemistry were trained primarily on-the-job. But by the 1920's, the value of professionals with extensive formal training had greatly increased. In the case of engineering and other applied sciences, the technological shock can be thought of as the use of formal science in industry. In the case of medicine, it can be thought of as the application of the scientific method. Because public institutions were set up, in part, to produce goods and services of value to the state's citizens and because these often took the form of research, public-sector

³ Joseph E. Hight (1975) computes a similar statistic for 1927-1972, but because his data do not contain the previous 30 years, he emphasizes the relative increase of the public sector after 1947 rather than seeing a greater increase before.

institutions could produce various types of training at lower cost than could most private institutions. Thus public institutions expanded their production of technical students relative to liberal-arts students and thereby expanded their share of all students relative to the private sector.

Even though private institutions were not all research institutions, they had something else to sell to prospective students. Some of the new professions, like business, and those under suspicion, like medicine, needed the reputational quality of the older private institutions. Thus some private institutions also expanded after the "technological shock."

Various facts are consistent with our story of the relative expansion of public-sector institutions from 1890 to 1940. No private institution of note was founded after 1900, whereas several had been in the 1890's (e.g., Stanford, the University of Chicago, California Institute of Technology). Something had changed around the turn of this century that erected a barrier to the opening of new institutions of higher education, particularly private ones. That barrier, we speculate, was the large size, reputation, and research resources required to be competitive with the state universities and with some of the older and more established private institutions. One piece of evidence that is consistent with the first hypothesis, but not with the second, is that expenditures per student increased more in the public than in the private sector from 1897 to 1923 (whereas expenditures were equal in 1897, the public-to-private ratio was 1.5 in 1923).⁴

An alternative, complementary, explanation for the rise in the public share of higher-education enrollment concerns the high-school movement that swept much of the nation between 1910 and 1940. Even though a smaller percentage of secondary-school graduates continued to college in 1925 than in 1905, more of those who did demanded the practical, applied, and scientifically oriented programs offered by state universities. Also, more of the newer graduates were less able to

afford private-sector tuition. Even though the explanation is a compelling one, the cross-state correlation of the high-school graduation rate in 1928 and the public share of higher-education students (among state residents) in 1931 is weak ($\rho = 0.17$). We can explain about half of the total change in the public share of higher-education enrollment from 1897 to 1940 on the basis of the cross-section relationship, leaving substantial room for the alternative hypothesis.

IV. What Explains Public Support for Higher Education?

State and local support to public higher education doubled from 1902 to 1940 as a fraction of all state and local expenditures (U.S. Department of Commerce, 1975 [series Y 684-85]), and the fraction of all students attending public-sector institutions also doubled over the same period. But public funding for higher education and access to public colleges and universities varied substantially among states throughout the period. The greatest levels of support were in the West, and the lowest were in the Northeast. What explains the differences among states and regions?

The determinants of the log of state and local per capita higher-education spending in 1929 are explored in the regressions of Table 1. The public-choice decision to provide support for higher education is likely to be affected by the level and distribution of wealth or income in a state, community stability and homogeneity, and the importance of industries that capture localized benefits of state institutions. Column (i) indicates a strong positive relationship between automobile registrations per capita and state support for higher education. A one-standard-deviation increase in this variable (0.32) is associated with a 0.4-log-point (49-percent) increase in state spending per capita on higher education. Auto registrations per capita in this period is a summary measure of both the level and distribution of wealth, since it is essentially a count of the fraction of individuals wealthy enough to own a car. Thus it represents the share of voters sufficiently wealthy to believe their children could attend college. The shares of employment in mining, manufac-

⁴ We compute these differences in a regression that includes controls for region and institution type (see Goldin and Katz, 1998).

TABLE 1—DETERMINANTS OF STATE SUPPORT FOR HIGHER EDUCATION, 1929

Independent variable	Regression		
	(i)	(ii)	(iii)
Log automobile registrations per capita, 1930	1.306 (0.278)	1.06 (0.274)	
Log agricultural income per agricultural worker, 1900			0.339 (0.153)
Fraction Catholic, 1910, 1926 ^a	-0.631 (0.584)	-0.628 (0.542)	-1.09 (0.515)
Fraction of labor force in mining, 1930	4.14 (1.59)	2.38 (1.62)	
Fraction of labor force in manufacturing, 1930	2.47 (1.57)	3.05 (1.47)	
Fraction of labor force in agriculture, 1930	1.73 (0.848)	1.45 (0.793)	
West (West North Central, Mountain, and Pacific)	0.803 (0.261)	0.782 (0.243)	
South (South Atlantic, East South Central, West South Central)	0.753 (0.244)	0.667 (0.229)	
East North Central	0.493 (0.206)	0.386 (0.195)	
Private college enrollments per 1,000 residents, 1900		-0.258 (0.0952)	-0.294 (0.115)
Year of statehood $\times 10^{-2}$			0.503 (0.202)
Constant	-1.68 (1.79)	-0.115 (1.76)	-3.88 (3.43)
R^2 :	0.759	0.798	0.645
Mean squared error:	0.322	0.298	0.371
Number of observations:	48	48	48

Notes: The dependent variable in all regressions is the log of per capita revenues of higher-education institutions from state and local governments. Numbers in parentheses are standard errors.

^a The 1926 figure is used for columns (i) and (ii); the 1910 figure is used for column (iii).

turing, and agriculture are also positively related to state support for higher education. These sectors may have lobbied effectively for research support at state institutions, or state legislators may have believed that the social return on public expenditures was high in a state with concentrations of agricultural products, unique manufacturing outputs, or

idiosyncratic engineering needs. The states of the Northeast (the base group) have far lower public support for higher education than do other regions, even including controls for wealth, industrial structure, and social makeup (fraction Catholic).

The regression in column (ii) of Table 1 adds enrollments in privately controlled institutions in 1900 as a proxy for the historical importance of private colleges and universities in each state. A substantial presence of private universities in a state in 1900 has a significant and depressing effect on public support of higher education in the state in 1929, but the inclusion of the variable does not markedly alter the impact of the others. The difference between private-college enrollments per 1,000 residents in Massachusetts and Iowa in 1900 (3.35 vs. 0.99) implies a decrease in per capita spending on higher education of 61 log points (84 percent).

In column (iii), we more fully explore, how state "initial" conditions, circa 1900, affected the expansion of public support for higher education (as measured in 1929). States with high agricultural income per worker at the turn of the century, a low share of Catholics, more recent statehood, and initially weak private universities provided more public support for higher education in 1929. We have also found a negative effect on public-college enrollments as a share of state population in 1929, from a strong early presence of private colleges (Goldin and Katz, 1998). Furthermore, greater access to public colleges, measured by lower tuition, is associated with a higher overall college enrollment rate (public and private) of state residents in 1929, even conditioning on per capita wealth and the high-school graduation rate. Thus, access to public higher education did not just affect enrollments in the publicly controlled sector. It also increased college matriculation overall.

Thus newer states, with a high share of well-to-do families and scant presence of private universities in 1900, became the leaders in public higher education by 1929 and remain so today. The tradition of stronger private universities and lower support for publicly controlled universities in the Northeast also continues to the present.

V. Concluding Remarks

The public sector in higher education rose relative to the private sector from the late 1890's to 1940. All institutions of higher education grew in scale and scope during those 50 years, and these changes have been associated with increased demand for professional training and enhanced complementarities between research and teaching. Public-sector universities were well placed to take advantage of these changes. The expanded pool of students produced by the high-school movement increased the demand for higher education among middle-class Americans. States more recently admitted to the Union and those with a lower preexisting presence of privately controlled institutions had greater public-sector expansions and became leaders in support for higher education.

The major state-level factors we identify that encouraged support to public higher education are a high level of wealth broadly distributed, the presence of business and commercial interests having large, unified, and concentrated demands for practical research, a late year of statehood, and a low early presence of private institutions. Differences across states in the relative role of the public sector in higher education that emerged during 1890–1940 persist today, despite the increased national scope of the market for higher education. But even though state rankings today are similar to what they were in 1930, the proportional differences in state spending on higher education and public enrollment per capita were far greater in the past.

REFERENCES

- Goldin, Claudia. "America's Graduation from High School: The Evolution and Spread of Secondary Schooling in the Twentieth Century." *Journal of Economic History*, 1998 (forthcoming).
- Goldin, Claudia and F. Katz, Lawrence. "Why the United States Led in Education: Lessons from Secondary School Expansion, 1910 to 1940," National Bureau of Economic Research (Cambridge, MA) Working Paper No. 6144, August 1997.
- . "Public and Private Higher Education: An Exploratory Study, 1890 to 1940." Working paper, National Bureau of Economic Research, Cambridge, MA, 1998 (forthcoming).
- Hight, Joseph E. "The Demand for Higher Education in the U.S., 1927–72; The Public and Private Institutions." *Journal of Human Resources*, Fall 1975, 10(4), pp. 512–20.
- Hoxby, Caroline M. "The Changing Market Structure of U.S. Higher Education," Working paper, Harvard University, 1997.
- Oleson, Alexandra and Voss, John, eds. *The organization of knowledge in modern America, 1860–1920*. Baltimore, MD: Johns Hopkins University Press, 1979.
- Quigley, John M. and Rubinfeld, Daniel L. "Public Choices in Public Higher Education," in Charles T. Clotfelter and Michael Rothschild, eds., *Studies of supply and demand for higher education*. Chicago: University of Chicago Press, 1993, pp. 243–78.
- Rosenberg, Nathan and Nelson, Richard R. "American Universities and Technical Advance in Industry," CEPR Publication No. 342, Stanford University, 1993.
- U.S. Department of Commerce, Bureau of the Census. *Historical statistics of the United States from Colonial times to 1970*. Washington, DC: U.S. Government Printing Office, 1975.
- U.S. Department of Education, National Center for Education Statistics. *Digest of education statistics 1996*. Washington, DC: U.S. Government Printing Office, 1996.
- Veysey, Laurence R. *The emergence of the American university*. Chicago: University of Chicago Press, 1965.

How Much Does School Spending Depend on Family Income? The Historical Origins of the Current School Finance Dilemma

By CAROLINE M. HOXBY*

Financing for U.S. elementary and secondary schools is in the throes of two crises, both of which are likely to affect the economy over the long term. The first is a crisis over the relationship between family income and per-pupil spending: to what degree should parents' income determine the amount of money spent on a child's elementary and secondary education? The answer is not only important for reasons of distributive justice. It may also be important for macroeconomic growth because it is not generally efficient to base human-capital investment in a child on the income of his parents (Michele Boldrin, 1993; Roland Benabou, 1996; Raquel Fernandez and Richard Rogerson, 1996). This question lies at the heart of cases filed over the constitutionality of states' school-finance systems (see Thomas Downes, 1992; Hoxby, 1995; William Evans et al., 1997).

The second crisis in school finance is over the property tax, which is the traditional and dominant source of revenue for public elementary and secondary education. The property tax can have good economic properties under optimal conditions (functioning as a user fee), but political dissatisfaction with the property tax is rising. Michigan largely eliminated it as the basis for school finance in 1994.

This paper explores the origins of these school-finance problems. Using data from 1900 to 1990, I investigate three questions. Has the distribution of per-pupil spending in the United States grown more or less unequal over time? Has the relationship between per-pupil spending and property value changed? Finally, has the relationship between parents' income and per-pupil spending become stronger or weaker over time? Underlying the questions is a classic puzzle about revolutions:

are the current crises due to increasing failure of the system or to rising expectations about what a school-finance system should be able to achieve?

I. Empirical Strategy

The questions posed require district-level data and a historical view. Data that fulfil these requirements must be gathered in a painstaking way from state archives because, until 1970, the federal government only gathered district-level data for large cities. This paper employs district-level data for Massachusetts, Illinois, and California—states chosen for data quality and representativeness (unfortunately, no Southern state could be included). School districts' records provided the data on expenditures, number of pupils, and local equalized property valuation. The U.S. Census of Population, aggregated to the same jurisdictional level as the school district whenever possible, provided the demographic data. (For pre-1970 censuses, small rural districts can sometimes only be matched to the rural portion of their county). Demographics include household incomes and the age distribution of the population. The data are described in detail in Hoxby (1998).

For studying school finance, it is essential to focus on *fiscal* school districts, that is, districts that have significant autonomy in revenue-raising and expenditure. They are the appropriate units to match with data on property values and per capita incomes. Fiscal districts should not be confused with attendance districts, which do not independently raise taxes or determine expenditures.

Each state started the century with a system based largely on local property taxes, although California's fiscal districts have always been substantially larger than those of Massachusetts and Illinois. Over time, all three states have given more control over school finance to the state government, but the state

* Department of Economics, Harvard University, Cambridge, MA 02138, and National Bureau of Economic Research.

governments have continued to rely on the property tax. By 1970, the states were using forms of foundation aid, which puts a floor under per-pupil spending by redistributing funds from districts with high property valuation per pupil to districts with low property valuation per pupil. Since then, Illinois and Massachusetts have done an increasing amount of redistribution, but maintained some local autonomy. The second Serrano decision caused California to move to statewide school finance during 1978–1980, so that the state now has one fiscal district allocating funds on a strict per-pupil basis over the vestigial (attendance) districts. Analyzing California is useful because, of the 50 states, it has had the most dramatic crisis and change in school finance. Thus, one might expect it to show the factors that generate a crisis.

II. Results

Table 1 presents a measure of inequality in per-pupil spending, the enrollment-weighted coefficient of variation.¹ The table shows that inequality in per-pupil spending was relatively stable from 1900 to 1990. The coefficient of variation is typically about 0.2, meaning that a standard deviation in per-pupil spending is about 20 percent of the mean. The student who experiences per-pupil spending at the 95th percentile goes to a school that spends about 60 percent more than the student at the 5th percentile. The difference between the 90th and 10th percentiles is about 45 percent, and the difference between the 75th and 25th percentiles is about 30 percent. Whether this amount of inequality is small or large depends on whether the standard of comparison is absolute equality or income inequality (which is much larger, as will be seen).

None of the measures of inequality varies dramatically over the century: the difference

TABLE 1—PER-PUPIL SPENDING INEQUALITY AMONG DISTRICTS, MEASURED BY THE COEFFICIENT OF VARIATION

Year	MA	IL	CA
1900	0.21	0.27	0.17
1910	0.21	0.26	0.18
1920	0.18	0.22	0.17
1930	0.18	0.21	0.16
1940	0.25	0.28	0.18
1950	0.16	0.21	0.15
1960	NA	NA	NA
1970	0.18	0.25	0.16
1980	0.24	0.28	—
1990	0.25	0.28	—

between the maximum and minimum coefficients of variation in per-pupil spending is about 8 percent. Nevertheless, there is a definite time pattern. Spending inequality was stable between 1900 and 1910, fell to a slightly lower level in 1920 and 1930, and widened again during the Great Depression. Inequality bottomed out in 1950, rose slightly between then and 1970, rose significantly between 1970 and 1980, and was relatively stable during the 1980's. Thus, if one looks at data from 1900 to 1950, spending inequality mainly appears to be falling. The reverse is true after 1950. Whether 1950 is a good benchmark depends on the purpose of the analysis. If we look at the entire century, spending inequality is surprisingly stable: current spending inequality is similar to that of 1900, 1910, 1940, and 1980.

There is an obvious parallel between the time pattern in spending inequality and that of economy-wide income inequality. In particular, the two decades in which spending inequality rose the most (the 1930's and the 1970's) were also decades of rising income inequality. The decade in which spending inequality fell the most (the 1940's) was a decade of falling income inequality.

Massachusetts, Illinois, and California all exhibit similar time patterns, but Illinois began with higher spending inequality than Massachusetts. California consistently had the lowest and least fluctuating spending inequality, probably because its large fiscal districts included a wider array of occupations and industries. The surprise is that California's

¹ Hoxby (1998) shows that similar patterns are obtained if Gini coefficients, log percentile spending ratios, or other measures of inequality are used instead of the coefficient of variation. Data for 1960 were not available. Weighting by enrollment is important for making comparisons over time because some districts have consolidated, and population densities have shifted. The Massachusetts data are consistently the highest in quality.

relatively stable and uniform per-pupil spending was more severely criticized during the decade from 1966 to 1976 than that of other states (certainly more than that of Massachusetts or that of Illinois). Of the three states, California apparently offered the least cause for complaint prior to its dramatic shift to state-level school finance.

Table 2 shows inequality of per-pupil valuation, using the enrollment-weighted coefficient of variation. Equalized property valuation (the valuation the state uses to calculate aid) is used. Per-pupil valuation is much less equal than per-pupil spending. A standard deviation in per-pupil valuation was between 55 percent and 65 percent of the mean from 1900 to 1990. Unlike inequality in per-pupil spending, there has been a long downward trend in the inequality of per-pupil valuation. The same fluctuations (more inequality in 1940, 1980, and 1990; less inequality in 1950) surround the negative trend. The downward trend is most obvious in Illinois, where the coefficient of variation fell from 65 percent in 1900 to about 45 percent during 1980–1990.

What explains the downward trend? The most likely explanation is the decline in arbitrary differences in districts' per-pupil valuation due to lumpy real-estate assets (like family-owned farms with few children residing on them). Over time, taxation of these assets has become increasingly distinguished from taxation of house property, both because of tax-law changes and because these assets have become corporate property. Districts have also consolidated, spreading these assets over a larger number of children.

The key implication of Table 2 is that the relationship between per-pupil spending and per-pupil valuation is not rigid. If it were rigid (if local residents did not modify their choice of property-tax rates depending on the relationship of local real-estate assets to incomes), then inequality in per-pupil spending would have fallen notably between 1900 and 1990.

The next step is to net out the effects of economy-wide income inequality on the inequality of per-pupil spending and valuation. This will allow me to focus on whether school-finance systems are operating differently over time or whether the systems are merely operating similarly in different environments. If the

TABLE 2—PER-PUPIL VALUATION INEQUALITY AMONG DISTRICTS, MEASURED BY THE COEFFICIENT OF VARIATION

Year	MA	IL	CA
1900	0.60	0.65	0.50
1910	0.55	0.62	0.48
1920	0.41	0.58	0.41
1930	0.41	0.57	0.40
1940	0.38	0.49	0.42
1950	0.37	0.41	0.36
1960	NA	NA	NA
1970	0.45	0.41	0.36
1980	0.48	0.45	0.39
1990	0.51	0.46	0.40

coefficients of variation in Tables 1 and 2 are adjusted by dividing by the overall coefficient of variation in income (income inequality in a state and year, based on the entire population and taking no account of school districts or pupils), the fluctuations in the inequality of spending and valuation per pupil are greatly dampened.² For Massachusetts from 1900 to 1990, the adjusted per-pupil spending series is nearly flat, and 1950 is not the nadir: 0.11 (1900), 0.12 (1910), 0.10 (1920), 0.12 (1940), 0.11 (1950), 0.11 (1970), 0.12 (1980), and 0.12 (1990). Adjusting per-pupil valuation makes the long-term fall in the inequality of per-pupil valuation more obvious. The series for Illinois is: 0.33 (1900), 0.35 (1910), 0.32 (1920), 0.25 (1940), 0.25 (1950), 0.24 (1970), 0.24 (1980), and 0.24 (1990).

Replication of Tables 1 and 2 for income per capita, (as opposed to spending and valuation per pupil) can only be done in a completely parallel manner for 1970 to 1990, because the availability of household-income data on the district level varies prior to these years. Table 3 shows the results of this exercise. The upper part of the table shows that between-district income inequality rose

² The series used for the denominator were calculated using Integrated Public Use Micro Sample data, which are available for every Census year in this century except 1930. Incomes for 1900, 1910, and 1920 were estimated from regressions of 1940–1960 wage and salary income to indexes of occupational status.

TABLE 3—PER CAPITA INCOME INEQUALITY AMONG DISTRICTS, MEASURED BY THE COEFFICIENT OF VARIATION

Year	MA	IL	CA
<i>A. Unadjusted</i>			
1970	0.22	0.29	0.21
1980	0.24	0.31	0.25
1990	0.29	0.33	0.28
<i>B. Adjusted for Economy-Wide Income Inequality</i>			
1970	0.13	0.16	0.14
1980	0.13	0.16	0.15
1990	0.14	0.16	0.15

between 1970 and 1990, while the lower part of the table shows that dividing by overall state-year income inequality eliminates this pattern. In other words, since 1970, there has not been increased sorting of households into districts based on income.

Table 4 uses regression to show how per-pupil valuation and per capita income explain per-pupil spending for Massachusetts and to see whether the relationships have changed over time. Estimates of β_1 and β_2 from the following regression are presented:

$$PPS_{ijt} = \beta_0 + \beta_1 PPV_{ijt} + \beta_2 PCI_{ijt} + \beta_3 OLD_{ijt} + \beta_4 HS_{ijt} + \beta_5 GRAD_{ijt} + \varepsilon_{ijt}$$

where PPS is per-pupil spending, PPV is per-pupil valuation, PCI is per capita income, OLD is the percentage of the population over age 65, HS is the percentage of high-school-aged children in the population, GRAD is the percentage of adults who are high-school graduates, and ε is an error term; i indexes districts, j indexes states, and t indexes time. The regression is run separately for each state and year.³ No regressions were run for California

³ Per capita income and the percentage of adults who are high-school graduates were estimated for 1900–1930 from regressions on indexes of occupational status and one-digit industry indicator variables. Varying the measure of income (median income or per-pupil income instead of per capita income) did not affect the regression

TABLE 4—ESTIMATED COEFFICIENTS FROM REGRESSION OF PER-PUPIL SPENDING ON PER-PUPIL VALUATION (PPV), PER CAPITA INCOME (PCI), PERCENTAGE ELDERLY, PERCENTAGE SCHOOL-AGE, AND PERCENTAGE HIGH-SCHOOL GRADUATES, MASSACHUSETTS

Year	Estimated coefficient	
	PPV	PCI
1990	0.17 (0.06)	0.35 (0.02)
1910	0.16 (0.05)	0.29 (0.02)
1920	0.19 (0.06)	0.33 (0.02)
1930	NA	NA
1940	0.19 (0.05)	0.32 (0.03)
1950	0.21 (0.04)	0.39 (0.02)
1960	NA	NA
1970	0.22 (0.03)	0.28 (0.03)
1980	0.24 (0.02)	0.20 (0.04)
1990	0.33 (0.02)	0.06 (0.04)

Note: Standard errors are given in parentheses.

in 1980 and 1990 because the state effectively had only one fiscal district.

Space constraints prevent the inclusion of estimates of β_3 , β_4 , and β_5 , but they may be summarized as follows. The percentage of the population over age 65 has a small positive, statistically significant (at the 5-percent level) effect on per-pupil spending in 1900 and 1910 (an additional 1 percent of the population being over 65 raises school spending by 0.5 percent). Estimates of this coefficient gradually reverse sign so that, by 1990, the percentage of the population over age 65 has a small *negative*, statistically significant effect on per-pupil spending (an additional 1 percent of the population being over age 65 lowers school spending by 0.7 percent). These results confirm those of Claudia Goldin and Lawrence Katz (1997) for the early part of the century and those of James Poterba (1997) for recent years. The percentage of households with school-age children has a consistently negative sign, and the percentage of adults who are high-school graduates has a positive, statistically significant effect on per-pupil spending until 1950 (but no statistically significant effect after that).

results. The results were also relatively insensitive to the inclusion of available demographic variables other than the three shown.

Table 4 demonstrates that per-pupil valuation and per capita income are powerful determinants of per-pupil spending. The equation explains the majority or a substantial minority of the variation in per-pupil spending among districts in each year. The R^2 coefficients for the Massachusetts regressions range between 0.31 (1940) and 0.58 (1950).

The first interesting pattern in Table 4 is that, from 1900 to 1970, per capita income is consistently a more powerful determinant of per-pupil spending than is per-pupil valuation. After 1970, the explanatory power of per capita income falls. Also, the estimated elasticity of spending with respect to per capita income falls from about 0.35 in 1900 to 0.06 in 1990 for Massachusetts and to 0.12 in 1990 for Illinois. The peak year for both explanatory power of per capita income and the elasticity of spending with respect to income is 1950 (for Massachusetts, the estimated coefficient is 0.39, and the associated t statistic is 18.0).

There are two likely explanations for this result. The first is that over the past 20 years, grants to districts where low-income households reside (such as Title I, special-education aid, and bilingual-education aid) have grown so much that low per capita income now has an ambiguous effect on per-pupil spending. The second possible explanation is that per-pupil valuation is increasingly an indicator of the local demand for per-pupil spending.

This second explanation is related to the other interesting pattern in Table 4. The explanatory power of per-pupil valuation has grown significantly over the century, and the estimated elasticity of per-pupil spending with respect to per-pupil valuation has increased. The increased explanatory power of per-pupil valuation is more surprising than the changing explanatory power of per capita income because nearly all school-finance reforms since 1970 (foundation aid, power equalization, guaranteed tax revenue) have tried to eliminate the influence of local per-pupil valuation on per-pupil spending. The results suggest that per-pupil valuation is increasingly an indicator for those elements of taste and household income that determine demand for per-pupil spending.

III. Conclusions

Per-pupil spending inequality has been relatively stable over the century, though there was a pronounced nadir of inequality in 1950. Most of the fluctuations in spending inequality were due to economy-wide changes in the inequality of household incomes. Spending inequality has not followed the pattern of inequality in per-pupil valuation, which has trended downward over the century.

Regressions of per-pupil spending on per-pupil valuation, per capita income, and demographic variables show that the relationship between spending and income grew stronger, in terms of both explanatory power and the elasticity of spending out of income, from 1900 to 1950 and then weakened. By 1990, there was no statistically significant relationship between local per capita income and per-pupil spending in the state of Massachusetts. The change was less dramatic for Illinois. The explanatory power of per-pupil valuation grew steadily from 1900 to 1990, despite school-finance equalization programs with redistribution formulas based (inversely) on per-pupil valuation. This is probably because per-pupil valuation has increasingly become an indicator for the locally preferred level of per-pupil spending.

Based on California's patterns of spending inequality, it would have been difficult to foresee that the state would have a school-finance crisis in the 1970's and ultimately abandon local school finance altogether. Of the three states, California had the lowest level and smallest fluctuations in inequality from 1900 to 1970. This evidence suggests that school-finance crises are not necessarily the result of systems breaking down, but may be due to rising expectations about the equality a state's school-finance system should be able to achieve.

The results have two implications for economics. First, the prevailing view of the relationship between income and spending may be too simplistic. Much of the empirical relationship between income and spending has already been eliminated. To reduce spending inequality, it would probably be more practical to focus on reducing income inequality than to redistribute an increasing share of existing

public-school revenues. Per-pupil valuation *does* matter for per-pupil spending, but probably not because of arbitrary differences in real property assets (which have apparently not had as much effect as claimed). Rather, per-pupil valuation is an increasingly good measure of local demand for school spending and should be viewed as such. Recent school-finance reforms treat per-pupil valuation as an asset impervious to the system of school finance. Future school-finance reforms should incorporate a more sophisticated understanding of valuation.

REFERENCES

- Benabou, Roland.** "Heterogeneity, Stratification, and Growth: Macroeconomic Implications of Community Structure and School Finance." *American Economic Review*, June 1996, 86(3), pp. 584-609.
- Boldrin, Michele.** "Public Education and Capital Accumulation," Northwestern University Economic Theory Workshop Discussion Paper No. 1017, 1993.
- Downes, Thomas.** "Evaluating the Impact of School Finance Reform on the Provision of Public Education: The California Case." *National Tax Journal*, December 1992, 45(4), pp. 405-20.
- Evans, William; Murray, Sheila and Schwab, Robert.** "Schoolhouses, Courthouses, and Statehouses After Serrano." *Journal of Policy Analysis and Management*, Winter 1997, 16(1), pp. 10-37.
- Fernandez, Raquel and Rogerson, Richard.** "Income Distribution, Communities, and the Quality of Public Education." *Quarterly Journal of Economics*, February 1996, 111(1), pp. 135-64.
- Goldin, Claudia and Katz, Lawrence.** "Why the United States Led in Education: Lessons from Secondary School Expansion, 1910 to 1940." National Bureau of Economic Research (Cambridge, MA) Working Paper No. 6144, August 1997.
- Hoxby, Caroline.** "Not All School Finance Equalizations Are Created Equal." Mimeo, Harvard University, 1995.
- . "The Empirical Relationship between School Expenditure, Income, and Property Wealth: Evidence for Macroeconomics Growth Models and Court Cases Over School Finance." Mimeo, Harvard University, 1998.
- Poterba, James.** "Demographic Structure and the Political Economy of Public Education." *Journal of Policy Analysis and Management*, Winter 1997, 16(1), pp. 48-66.

Demographic Change, Intergenerational Linkages, and Public Education

By JAMES M. POTERBA*

The prospective changes in population-age structure that have stimulated discussion of the economic viability of transfer programs that benefit the elderly, such as Social Security and Medicare, also raise questions about the political viability of programs that transfer resources to children. If voters decide which expenditure programs to support on the basis of narrowly defined self-interest, as at least some politico-economic models assume they do, then as a society ages, there may be diminished support for youth-targeted programs such as public education. Samuel Preston (1984) argues that the growing political influence of elderly voters, and more generally of voters from childless households, may have important effects on the pattern of age-specific government transfer programs.

In this brief paper, I explore several issues related to demographic change and the political economy of public education. I begin with a brief summary of the projected demographic changes that will take place in the United States over the next three decades. Next, I describe the existing empirical evidence that suggests that older and childless voters are less likely to support public-school spending than younger voters with children. I then note several unresolved issues about the degree to which rational self-interest should lead older voters to vote for low levels of public-school spending. The closing section speculates about whether there have been changes in the age-specific patterns of political support for public education, and if so, what might account for such changes.

I. Demographic Change: Young and Old Dependents

The age distribution of the U.S. population will change significantly in the next few decades. In 1996, 12.6 percent of the population was over the age of 65, while 25.7 percent was below the age of 18. Census Bureau projections suggest that, by 2030, the population share over the age of 65 will rise to 20 percent, while the share under the age of 18 will decline to 24 percent. By 2030 the share of over-65 individuals in the aggregate U.S. population will exceed that in Florida today. The ratio of elderly individuals to those under the age of 18 will rise from 0.49 in 1996 to 0.83 in 2030.

The declining share of children in the overall population will coincide with a rising share of children from minority groups. In 1995, 20 percent of those under the age of 18 were non-whites. By 2025, Census projections suggest that this share will rise to 26 percent. The decline in the share of children in the population will also coincide, not surprisingly, with a decline in the fraction of households with school-age children. In 1996, 34.3 percent of all households were family households with children under the age of 18 present. By 2010, Census projections suggest that this fraction will decline to 28 percent. In 1960, when the baby-boom cohort was of school age, 48.7 percent of all households were family households with children under 18. The prevalence of such family households is an important factor in analyzing the political economy of public education, because previous empirical work suggests that an individual's political support for public education depends strongly on whether he lives in such a household.

Incipient demographic changes will shift the age composition of dependent individuals in the population more than they will shift the total share of the population that is dependent. The elderly support ratio, the number of individuals aged 65+ divided by the number aged

* Department of Economics, Massachusetts Institute of Technology, 50 Memorial Drive, Cambridge, MA 02142-1347. I am grateful to Claudia Goldin, Caroline Hoxby, Robert Inman, Christopher Jencks, Helen Ladd, Lawrence Katz, and Douglas Wolf for helpful comments, and to the National Institute of Aging and the National Science Foundation for research support.

18–64, will rise from 20 percent in 1996 to 38 percent in 2030. The total support ratio, however, the number of persons under the age of 18 plus the number aged 65+, all divided by the number aged 18–64, will rise from 61 percent to 74 percent. These total support ratios are lower than those experienced in the first two-thirds of the present century. In 1900, the total support ratio was 84 percent (76 percent from children). In 1960, this ratio was 82 percent (65 percent from children), and in 2010, it is projected to reach its lowest level in more than a century at 57 percent (35 percent from children).

II. Is Support for Public Education Age-Dependent?

Concern that an aging population will not support spending on public education is premised on the assumption that older voters are less likely than younger voters to support tax-financed spending on public schools. A growing body of empirical evidence suggests that this is the case. Three strands of evidence warrant discussion.

First, public-opinion surveys typically find greater support for public-school spending among younger voters and those with school-age children. Maris Vinovskis's (1993) analysis of data from the 1988 American National Election Study finds that 77 percent of respondents between the ages of 18 and 29 supported additional federal assistance for public schools, compared with 47 percent of those aged 70 and above. Daniel Rubinfeld (1977) analyzes data from a household survey with information on preferences about the level of school spending. He finds that whether the household has children in the local public-school system has a substantial and positive effect on whether the household head supports higher spending on local public schools. The age of the household head does not have any additional explanatory power once children's use of the public schools is controlled for. James Wyckoff (1984) reports similar findings from another survey of voters in a Michigan town. Although Rubinfeld's findings are sometimes cited as suggesting that age is not an important determinant of school support, in an aging population the fraction of households

with school-age children typically declines. Thus Rubinfeld's results suggest that prospective demographic changes may reduce support for public-school spending.

Second, empirical analyses of the outcomes of referenda on school finance suggest that, in communities with a higher fraction of older residents, school bond issues are more likely to be defeated. James Button (1992) presents a detailed analysis of voting on school bond referenda in six Florida counties. He finds that, in five of the six counties, the percentage of voters over the age of 55 has a negative and statistically significant effect on the probability of a precinct's approving a school bond issue. This does not appear to be the result of general antitax sentiment on the part of elderly voters; there is no statistically significant relationship between the fraction of voters aged 55+ and the approval of tax increases that are not related to schools. Because elderly individuals are more likely to register and to vote than are younger persons, they may have an effect on electoral outcomes that is disproportionate to their population share.

"Tax-limitation" referenda outside Florida also suggest that elderly voters are more likely to support tax-limitation legislation that may constrain local school expenditures. Helen Ladd and Julie Boatright Wilson (1983) show that elderly voters were more likely than younger voters to support Massachusetts' Proposition 2½ referendum in 1980. They cite related evidence showing similar age patterns in support for a tax-limitation bill in Michigan, although they note that there is little evidence that older voters were more likely to support Proposition 13 in California.

Anecdotal evidence supports these statistical findings, and it further suggests that elderly voters are particularly opposed to spending on public education when the individuals who benefit are from ethnic groups other than their own. William Bulkeley (1991) describes the experience of Holyoke, Massachusetts, where elderly white voters do not support programs that benefit young nonwhites. As the composition of Holyoke's school-age population has shifted toward nonwhites, political support for the schools has diminished. This observation is consistent with Alberto Alesina et al.'s

(1997) finding that local spending on publicly provided goods, including education, is lower in jurisdictions with ethnic fragmentation than in homogeneous communities.

The third source of evidence on demographic structure and public-school spending is cross-sectional and panel-data studies of local or state spending on public education. Robert Inman (1978) presented evidence of this type for Long Island school districts. He found that a 1-percent increase in the fraction of households in a school district that were headed by someone over the age of 64 reduced school spending by roughly 0.3 percent. David Cutler et al. (1993) survey a number of additional studies of education spending at the school-district and more aggregated level, many of which find negative effects of the elderly share of the population on either education spending per capita or per pupil. A negative effect on per capita spending is not surprising: even if expenditures per pupil were independent of the demographic structure of the voting population, one would expect per capita spending to fall as the share of children in the population declined.

My own recent study, Poterba (1997), investigates the link between demographic structure and the state-wide average level of public-school spending per student using a panel data set on the continental U.S. states for the 1961–1991 period. The results suggest that the level of per-child education spending is substantially (and statistically significantly) lower in states in which a greater share of the population is over the age of 65. The three key coefficients from the central regression specification, in which the dependent variable is the natural logarithm of per-child school expenditures, are shown in Table 1.

These results suggest that per-pupil expenditures are lower in states with a higher fraction of older persons, and particularly in states with a substantial population of older individuals who are from a different ethnic or racial group than the school-aged population. Expenditures per pupil also appear lower in states with large school-age populations. This is consistent with incomplete adjustment of aggregate spending to fluctuations in the number of school-age children.

TABLE 1—KEY RESULTS OF REGRESSION
FROM POTERBA (1997)

Independent variable	Coefficient	SE
ln(population share aged 65+)	−0.244	0.122
ln(population share aged 5–17)	−1.025	0.212
Difference between nonwhite share of the population aged 5–17 and 65+	−0.621	0.394

These results are suggestive, but they should be viewed with caution. Claudia Goldin and Lawrence Katz (1997) find that, at the turn of the 20th century, there was apparently more support for expanding public high schools in states with more older voters; whether this is a proxy for other state conditions or a direct effect of demographic factors on education support is not clear. My (1997) findings with respect to ethnic differences between old and young residents are also sensitive to the inclusion or exclusion of other covariates, notably an indicator variable for the fraction of a state's population living in urban areas.

While mindful of these limitations, I tried to evaluate the potential effects of demographic change on public education as “predicted” by these equations. I computed the difference in predicted per-student education expenditures in two states, one with 12.6 percent of its population over the age of 65, and the other with 20 percent of its population in this age group. These hypothetical “states” have the elderly population shares corresponding to the United States in 1996 and 2030, respectively. The coefficient estimates presented in Table 1 suggest that per-pupil expenditures would be 12.2-percent lower in the “older” state. At 1997 expenditure levels, this would represent approximately a \$700 per-student expenditure decline. Evaluating how such a spending change would affect student performance and educational outputs is beyond the current study. It should be noted that similar calculations based solely on demographic factors would have predicted a decline in real per-pupil spending of 7.7 percent, even though actual spending rose over this period.

III. Reasons Why Self-Interested Elderly Voters May Support Public Education

The discussion so far has assumed that elderly voters vote based on their generational self-interest, and that this self-interest is best served by low levels of expenditures on public schools. In such a setting, as Antonio Rangel (1997) shows, the share of older voters may affect equilibrium expenditures, and there may be potential inefficiencies because future generations are unable to trade with current generations. The empirical evidence described above is broadly supportive of the assumptions that underlie these results. A number of objections can be raised, however, to each of these assumptions; this section outlines four of these arguments as a roadmap for future research.

First, intergenerational externalities may lead older voters to support educational spending even though it does not benefit them directly. Harold Richman and Matthew Stagner (1986) argue, in contrast to Preston (1984), that a rising number of elderly households could lead to greater government transfer flows toward the young as the elderly seek to raise the training of younger workers. Greater training for young workers would raise the pool of resources from which transfers to the elderly could be funded, and it would also raise the quality of services that the elderly receive from younger workers.

Second, it is possible that the assumption of self-interested generational voting is incorrect, or that intergenerational altruism overwhelms the elderly's opposition to education-supporting taxes. John Logan and Glenna Spitze (1995) dispute the basic premise that voters are interested primarily in their generation's welfare; they argue that altruism is a more appropriate model. David Stromberg (1997) presents a careful analysis of the difference between median-voter and social-planner outcomes in an economy in which voters are altruistically linked to both their parents and their children. The degree of altruism, both from older to younger individuals and vice versa, can be a critical determinant of the age-specific structure of government expenditures and taxes. Measuring the degree of altruism is difficult, however, and it is therefore difficult to evaluate this argument with any-

thing other than reduced-form statistical evidence.

A third circumstance in which older voters might support spending on education involves property-value capitalization effects. If potential home-buyers are concerned about school quality and are willing to pay more for better schools, then property-value maximization may lead older voters to support a higher level of school spending than they would otherwise. Property values are more likely to capitalize differences in school spending across jurisdictions within a metropolitan area, among which individuals may be making residential location choices, than across jurisdictions separated by longer distances. The capitalization argument also requires that elderly voters be able to borrow against the accumulation of value in their homes, or otherwise transform the gains from property appreciation into current consumption. The degree of property-value capitalization is likely to vary across places and also potentially over time.

A final factor that might break the link between the fraction of older individuals in the aggregate population and the level of per-pupil school spending is "Tiebout sorting" of individuals by tastes for such spending. If elderly households who do not wish to pay high taxes in support of public schools can move to communities with low levels of school spending, then the level of per-pupil spending in districts with large numbers of children could remain high even as the overall population ages.

Tiebout sorting seems unlikely to undo completely the effect of aggregate population aging on support for public education. The limited empirical evidence on migration decisions of the elderly provides a mixed message with respect to the effect of education spending and local taxes. Karen Conway and Andrew Houtenville (1998) find some evidence that the elderly are less likely to move to states with high per capita education expenditures. However, they also find evidence that the elderly are less likely to leave high-spending states. Moreover, migration within states offers a limited mechanism for sorting households into high-education-spenders and low-education-spenders. Centralized state funding of local public schools has become

more important in the last decade, as have court-ordered reductions in the disparities in spending per pupil across school districts. These developments make it more difficult for parents of school-age children who wish to spend heavily on public schools to segregate themselves into communities that will do so. They also make decisions about school-spending levels more dependent on state as opposed to local electorates and thereby weaken the power of community location to affect the provision of education.

The arguments developed above suggest that the link between population age structure and the level of public support for publicly provided education is ambiguous on theoretical grounds. Further empirical work is therefore needed to provide additional evidence on the net effect of demographic structure on the level of public-school spending.

IV. Conclusion and Historical Perspective

Limited historical evidence suggests that age-related differences in support for public education have not always been present in the United States. Caroline Hoxby (1997) presents fascinating evidence on the changing determinants of school spending in California, Illinois, and Massachusetts over the 1900–1990 period. She studies the links between per-pupil expenditures in individual school districts and the level of per capita income, per-pupil property-tax valuation, and district demographic structure. In all three states at the beginning of her sample (1900), she finds a statistically significant and positive effect of the elderly population share on school spending; this is consistent with Goldin and Katz's (1997) findings on high schools. By the end of her sample, however, Hoxby finds a statistically significant and negative effect of the same variable. There is a relatively smooth positive-to-negative pattern in the estimated coefficients for all three states. These cross-sectional findings represent important support for the panel-data findings on state expenditure levels described above.

These findings raise the critical question of whether the nature of political support for generation-specific publicly provided goods has changed over time, and if so, why this

change has taken place. One potential explanation is a decline in the family and other links that connect older and younger residents of communities. Eileen Crimmins and Dominique Ingegneri (1990) discuss the limited empirical evidence on the fraction of the elderly who live with or near their children. They present both historical and contemporary information, along with reference studies suggesting that rising rates of internal migration may weaken links between the elderly and their children. Census data do not show a recent increase in mobility rates among older households, although for this argument, mobility among younger individuals could also weaken intergenerational ties. In the 1960 Census, for example, 70 percent of those over the age of 65 reported that they lived in the same house that they lived in five years earlier. The analogous statistic in 1980 was 77 percent, and in the mid-1990's, the annual mobility rates for individuals over the age of 65, based on the Current Population Survey, were just over 5 percent per year. Data on the proximity of older individuals to younger family members is difficult to obtain, and there is the further possibility that nongeographic factors may also affect the strength of intergenerational linkages.

To provide a complete description of how the demographic changes that will affect the United States in the next half century will affect the level and composition of public spending, it is important to consider how an aging electorate may alter the nature of expenditure programs. This issue has not received as much attention as the impact of demographic change on the solvency and structure of Medicare and Social Security, and it deserves further analysis. This paper has outlined a number of issues that must be addressed in this research program.

REFERENCES

- Alesina, Alberto; Baqir, Reza and Easterly, William. "Public Goods and Ethnic Divisions." National Bureau of Economic Research (Cambridge, MA) Working Paper No. 6009, 1997.
- Bulkeley, William M. "Hard Lessons: As Schools Crumble, Holyoke, Mass. Voters

- Reject Tax Increases." *Wall Street Journal*, 25 November 1991, 218, pp. A1, A6.
- Button, James W. "A Sign of Generational Conflict: The Impact of Florida's Aging Voters on Local School and Tax Referenda." *Social Science Quarterly*, December 1992, 73(4), pp. 786-97.
- Conway, Karen and Houtenville, Andrew. "Do the Elderly 'Vote With Their Feet'?" *Public Choice*, 1998 (forthcoming).
- Crimmins, Eileen and Ingegneri, Dominique. "Interaction and Living Arrangements of Older Parents and Their Children." *Research on Aging*, March 1990, 12(1), pp. 3-35.
- Cutler, David M.; Elmendorf, Douglas W. and Zeckhauser, Richard J. "Demographic Characteristics and the Public Bundle." *Public Finance/Finances Publiques*, Supplement 1993, 48, pp. 178-98.
- Goldin, Claudia and Katz, Lawrence. "Why the United States Led in Education: Lessons from Secondary School Expansion, 1910-1940." National Bureau of Economic Research (Cambridge, MA) Working Paper No. 6144, 1997.
- Hoxby, Caroline M. "How Much Does School Spending Depend on Family Income? The Historical Origins of the Current School Finance Dilemma." Mimeo, Harvard University, 1997.
- Inman, Robert P. "Testing Political Economy's 'As If' Proposition: Is the Median Income Voter Really Decisive?" *Public Choice*, 1978, 33(4), pp. 45-65.
- Ladd, Helen F. and Wilson, Julie Boatright. "Who Supports Tax Limitations? Evidence from Massachusetts' Proposition 2 $\frac{1}{2}$." *Journal of Policy Analysis and Management*, Winter 1983, 2(2), pp. 256-79.
- Logan, John R. and Spitze, Glenna. "Self Interest and Altruism in Intergenerational Relations." *Demography*, August 1995, 32(3), pp. 353-64.
- Poterba, James M. "Demographic Structure and the Political Economy of Public Education." *Journal of Policy Analysis and Management*, Winter 1997, 16(1), pp. 48-66.
- Preston, Samuel. "Children and the Elderly in the United States." *Demography*, November 1984, 21(4), pp. 435-57.
- Rangel, Antonio. "Intergenerational Goods." Mimeo, Harvard University, 1997.
- Richman, Harold A. and Stagner, Matthew W. "Children: Treasured Resource or Forgotten Minority?" in Alan Pifer and Lydia Bronte, eds., *Our aging society: Paradox and promise*. New York: Norton, 1986, pp. 161-79.
- Rubinfeld, Daniel L. "Voting in a Local School Election: A Micro Analysis." *Review of Economics and Statistics*, February 1977, 59(1), pp. 30-42.
- Stromberg, David. "Demography, Voting, and Local Public Expenditures: Theory and Evidence from Swedish Municipalities." Mimeo, Princeton University, 1997.
- Vinovskis, Maris. "An Historical Perspective on Support for Schooling by Different Age Cohorts," in Vern L. Bengtson and W. Andrew Achenbaum, eds., *The changing contract across generations*. New York: Aldine de Gruyter, 1993, pp. 45-65.
- Wyckoff, James H. "The Nonexcludable Publicness of Primary and Secondary Public Education." *Journal of Public Economics*, August 1984, 24(3), pp. 331-52.

WORK OR LEISURE: A CHANGING DECISION?[†]

When We Work

By DANIEL S. HAMERMESH*

Theories and empirical analyses of work and leisure have concentrated on constructing models and discovering facts about the *amount* of these activities, especially hours per day, per week, per year, or over a lifetime, and whether a person works at all in a particular week. With few exceptions (Gordon Winston, 1982; Hamermesh, 1996) *when* they are undertaken has not been considered. Analyzing and discovering facts about the timing of leisure are important for understanding labor markets, the distribution of economic welfare, and the nature of macroeconomic fluctuations.

I. The Importance of Timing

Individuals do not produce and consume commodities in vacuo. The utility-maximizing household faces a coordination problem in trying to maximize its well-being by producing the optimal amount of each commodity at the optimal times. Dinner or sex together with one's spouse is in most cases more appealing than the same activity undertaken alone; and not surprisingly, husbands and wives do in fact time their work to enjoy leisure together. (At each hour of the day in 1991 the correlation between the probabilities that one working spouse was on the job and that the other one also was is at least +0.10, significantly different from zero [Hamermesh, 1996 Ch. 3].) Similarly, in many businesses productivity depends on the presence of colleagues (consider

an assembly line) in the same plant and often in plants and firms located elsewhere.

These considerations yield positive predictions about timing and economic welfare over time. In addition to generating (possibly offsetting) income and substitution effects on the mix of goods- and time-intensive activities (Gary Becker, 1965), an increased value of time also generates effects on the timing of interactions with family members and others. To the extent that technical change produces temporally neutral changes in households' opportunities, household members in growing economies will be observed using some of their increased opportunities to schedule household production to allow greater coordination in consumption activities. The result of these changes in the timing of consumption activities is a welfare-improving change in the timing of work activities.

There are limits on coordinating the timing of consumption and work. Diurnal light-dark cycles impose circadian rhythms on most humans which generate biological responses leading to sleep at particular times and non-sleep activities at others (e.g., Jürgen Aschoff, 1982). Within these biological constraints, however, there is tremendous scope for cultural and historical differences to generate different timing outcomes (try finding dinner in Madrid before 9:00 P.M.). Beyond these noneconomic factors, changes in the price of time can induce changes in the timing of work and leisure even absent any desire for improved coordination, as individuals reoptimize their own well-being in the face of an expanded choice set.

The timing of leisure and work may also help shed light on macroeconomic fluctuations. In general-equilibrium models technical shocks generate fluctuations in leisure timing that feed back into the timing of market production. These in turn affect the time paths of

[†] *Discussants:* Casey B. Mulligan, University of Chicago; Shelly Lundberg, University of Washington; F. Thomas Juster, University of Michigan.

* Department of Economics, University of Texas, Austin, TX 78712, and NBER. I thank Jeremy Atack, Dora Costa, Robert Goldfarb, and Robert Margo for helpful comments, and the National Science Foundation for support under grant SBR-9422429.

market quantities, since output cannot be fully independent of when it is produced. Because the coordination of activities is almost definitionally a matter of when they occur rather than their quantities, understanding the micro foundations of the timing of leisure is essential for macroeconomics.

II. Historical and Contemporary Changes in Timing

The best way to study the timing of leisure and work is to analyze time budgets, data on individuals or entire households based (usually) on quarter-hour diaries kept for one or several days (e.g., Thomas Juster and Frank Stafford, 1985). Such data are unavailable for historical times, so one must rely on historians' observations based on textual information. The *Annales* School generated the series, *La Vie Quotidienne en . . .*, which allows us some inkling of the timing of activities in pre-industrial times. The overwhelming impression is that of the importance of the light-dark cycle absent reliable indoor lighting. As one of many examples in this series, in Periclean Greece, "The working day, and also every kind of public meeting, . . . began, in the normal way, at daybreak" (Robert Flacelière, 1965 p. 167). Reading these histories and comparing them to experiences during the Industrial Revolution and today also makes it clear that work has become more mixed with leisure: we have fewer holidays, but longer (and more frequent) weekends.

To understand work timing a century ago one can compare two tidbits of historical information to roughly similar tabulations from evidence on starting and ending times of work on the main job from the Current Population Surveys (CPS) of May 1985 and 1991. From Martha Shiells and Gavin Wright (1983 p. 343), the average fraction of workers on night shifts in the North Carolina textile industry between 1905 and 1926 was 0.270. The fraction of production workers in North Carolina textile plants working more than four hours between 8:00 P.M. and 6:00 A.M. during 1985–1991 was 0.231, which probably includes some workers who worked mostly in the evenings or even during the day. Less closely comparable are Boston "working

girls" in 1884 (Carroll Wright, 1889) and women in the Boston metropolitan statistical area in 1985–1991. The fraction of women who were free to determine their schedules and who worked on shifts that included hours between 8:00 P.M. and 6:00 A.M. in 1884 was 0.080.¹ The fraction working more than four hours between 8:00 P.M. and 6:00 A.M. during 1985–1991 was 0.058. The invention of bright artificial lighting, while initially allowing employers to extend daily work hours into the evening and night, helped create the economic growth that allowed greater choice in timing work and leisure.

There are no time series of time budgets for the United States. If one is willing to use short-term retrospective questions, comparable information can be obtained from large representative samples covering the second half of post-World War II U.S. history. From 1973 through 1978, and in 1985 and 1991, the May CPS included the questions about usual starting and ending times and asked workers how many days they worked. (Regrettably, only in the last two years did the CPS obtain information on which days they worked.) To examine trends I transformed the data from 1973 and 1991 into measures indicating whether an individual was working at a particular hour of the day.

Figure 1 shows the change between 1973 and 1991 in the fraction of workers at work at each hour. The base is all workers who report positive usual weekly hours (but the figure looks essentially identical if I restrict it to those working at least half-time, or to those working full-time). Two striking changes have occurred: (i) evening and night work became less important; and (ii) work at the fringes of the "normal" work day (6:00–7:00 A.M. and 5:00–6:00 P.M.) grew in importance. While some of these changes are small abso-

¹ Massachusetts banned women's work in textile manufacturing between 6:00 P.M. and 6:00 A.M., so I assume that no textile workers in the sample worked between 8:00 P.M. and 6:00 A.M. Assuming that other women in manufacturing, where the legislation prohibited work between 8:00 P.M. and 6:00 A.M., could be at work around 8:00 P.M. may impart a downward bias to this estimate of the propensity to work among women who were free to determine their schedules.

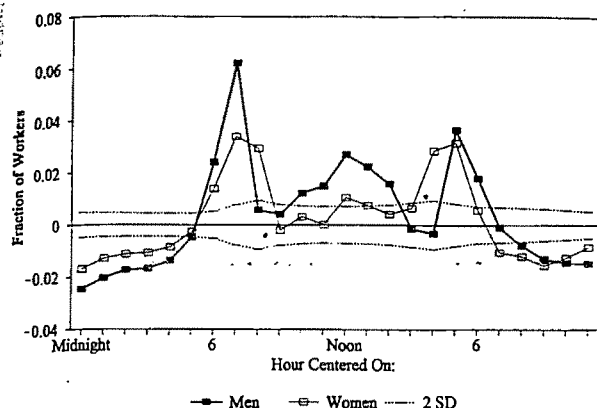


FIGURE 1. CHANGE IN FRACTION AT WORK, 1973-1991

lutely, they are large relatively. For example, the fraction of men at work at 3:00 A.M. in 1973 was only 0.084, so that the drop of 0.017 represented a 20-percent decline.²

If the timing of people's leisure and work activities responds to economic incentives, these highly significant (statistically and economically) changes are informative about changes in households' well-being. There is substantial evidence (for both the United States and Germany) that evening and night work are inferior (Hamermesh, 1996). That being the case, the trends depicted in Figure 1 are consistent with individuals' responses to higher real full incomes and imply that the welfare of the average worker increased over this period. The growing concentration of work at the periphery of the "normal" work day is harder to interpret in the context of consumer choice, absent temporally biased technical change, and may be due to growing costs of commuting during usual rush hours.

III. Changing Days of Work

Weekly work hours in the United States dropped sharply during the first 40 years of the 20th century, with a concomitant move away from Sunday, and then Saturday work.

² The figure is essentially unchanged if the sample is restricted to the 94 percent of workers who held only one job. Also, while there are cross-section differences among major industrial sectors, in nearly all of them the same changes have occurred over time.

Since then weekly and annual hours have changed little if at all (Mary Coleman and John Pencavel, 1993). At the same time the daily fixed costs of work (commuting costs, taking children to child-care, and others) may have changed relative to the per-hour costs (e.g., hourly child-care) in such a way as to alter the mix of days and daily hours. If commuting costs *for a trip taken at the same time of day* have risen, workers and the firms with which they implicitly contract have an incentive to work more hours on fewer days. By raising the supply price of days of work for a fixed total number of weekly work hours, these changes would have caused an increase in the fraction of workers on four-day or even shorter work weeks during the postwar period. Even if the relative supply prices of days and hours did not change, one would expect income effects and rising real full incomes to have shifted the distribution of work effort toward relatively hours-intensive schedules: there is substantial cross-section evidence (Hamermesh, 1996 Ch. 2) that concentrating a fixed-length work week on fewer days is superior.

Table 1 presents the distributions of days worked by men and women in 1973 and 1991. The distributions did change significantly: There was a decline in the fraction of workers with exactly five-day schedules, and there was an increase in the fraction with four-day schedules. Most important in light of this discussion, for both sexes the fraction working on fewer than five days increased, from 7.3 percent to 8.9 percent among men, from 17.6 percent to 18.4 percent among women. While these data suggest some movement in the direction of working on fewer days, the change is small indeed. The fraction working on at least six days declined slightly among men (from 21.2 percent to 20.5 percent), but it actually rose among women (from 8.5 percent to 8.7 percent). The data are only slightly suggestive of the impact of rising daily costs of supplying work or employing workers. Barring a resumption in the decline in weekly work hours or sharp increases in commuting costs, rumors of the demise of the five-day work week are greatly exaggerated.

TABLE 1—DISTRIBUTION OF DAYS WORKED PER WEEK, 1973 AND 1991

No. of days	Men		Women	
	1973	1991	1973	1991
1	1.4	0.5	3.0	1.3
2	1.5	1.3	4.2	3.3
3	2.0	2.5	5.5	6.3
4	2.4	4.6	4.9	7.5
4.5	0.4	0.7	0.7	1.0
5	71.1	69.9	73.2	71.9
5.5	5.6	4.0	1.8	1.7
6	13.0	12.1	5.6	5.0
7	2.6	4.4	1.1	2.0
p:	<0.001		<0.001	
N:	26,833	31,541	17,798	27,382

Note: Calculated from CPS, May 1973 and 1991.

IV. Interactions of Work Timing and Work Days

When people work and the particular days on which they work are related: as long as the worker's costs of working and the employer's labor costs at different times of the day vary across the week, both will change in response to changes in the price of days or of working at different hours, or to changes in full incomes. For example, a rise in commuting costs will increase workers' and employers' incentives to schedule work on fewer days and to shift the timing of work away from the most congested periods, even for a work week of fixed length.

Historical records give even less insight into these interactions than they do to work days or the timing of work separately. However, it is possible to examine the issue over the period 1973–1991 in the United States. Consider the series of probits:

$$(1) \Pr\{L_{it} = 1\} = \alpha_{0t} + \alpha_{1t} \text{DAYS}_i + \alpha_{2t} \text{HRSWK}_i + \varepsilon_{it} \quad t = 1, \dots, 24$$

where L_{it} indicates whether person i works at hour t , DAYS is days worked per week, HRSWK is total weekly work hours, ε is an error term, and the α 's are parameters. I estimate equation (1) separately by sex using the same May CPS data. By holding weekly work hours

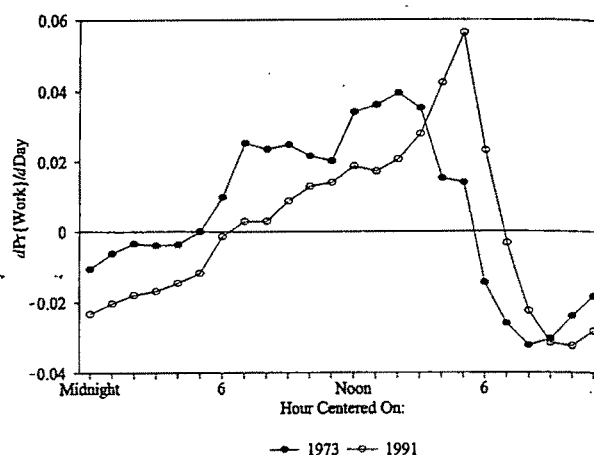


FIGURE 2. WORK TIMING AND WORK DAYS, MEN, 1973 AND 1991

constant, these equations focus on the relation between timing and work days. Thus the α_{1t} 's solely reflect interactions of workers' preferences and employers' costs independent of the quantity of labor supplied or the demand for total weekly hours.

Figure 2 presents the estimates of α_{1t} for men. (The figure for women looks qualitatively the same.) Nearly all the estimates are highly significantly nonzero. They demonstrate that working more days *within a given workweek* is associated with additional work during the main part of the business day. Conversely, those who work unusually many days are less likely than other workers to work evenings or nights. Adding the entire array of demographic and industrial/occupational variables available from the CPS does not materially change the figure. Most important, there is no significant change in this interaction between 1973 and 1991. Daily schedules and the pattern of days worked changed significantly over this period, but their interaction did not.

The results suggest that the unpleasantness of working at inferior times is partly offset by a less frequent need to incur the fixed daily costs of working. The case of the 12-hour-shift night nurse is an example. This interpretation suggests that the welfare implications of daily work timing may be mitigated by differences in workers' combinations of days and daily hours. The results present an anomaly, however: If evening and night work are

inferior, and working additional days for a given total weekly hours is inferior, how can it be that additional days are associated with additional work during the usual business day (i.e., at times that the evidence suggests are superior)?

V. Work Timing—Insufficient Study, Insufficient Data

The meager available data suggest that there have been important historical and contemporary changes in the timing of work and leisure over the day. Similarly, people's annual and weekly cycles of work have changed over time. Both sets of changes are amenable to economic analysis, but both have received remarkably little attention. While immense resources are concentrated on collecting information about the quantity of work and leisure in modern economies, very little effort is spent on obtaining data on their timing. The disparity in social scientists' attention to these two aspects of time use is even greater. Regrettably, without new data sets, particularly larger versions of the very few time-budget studies that exist for the United States, it is unlikely that much more will be learned about the determinants of work and leisure timing, how they have changed, and how they affect other economic outcomes.

REFERENCES

- Aschoff, Jürgen. "Circadian Rhythms in Man," in John Brady, ed., *Biological timekeeping*. Cambridge: Cambridge University Press, 1982, pp. 143–56.
- Becker, Gary S. "A Theory of the Allocation of Time." *Economic Journal*, September 1965, 75(3), pp. 492–517.
- Coleman, Mary T. and Pencavel, John H. "Trends in Market Work Behavior of Women Since 1940." *Industrial and Labor Relations Review*, January 1993, 46(3), pp. 653–76.
- Flacelière, Robert. *Daily life in Greece at the time of Pericles*. New York: Macmillan, 1965.
- Hamermesh, Daniel S. *Workdays, workhours, and work schedules: Evidence for the United States and Germany*. Kalamazoo, MI: W. E. Upjohn Institute, 1996.
- Juster, F. Thomas and Stafford, Frank P. *Time, goods, and well-being*. Ann Arbor, MI: University of Michigan Press, 1985.
- Shiells, Martha E. and Wright, Gavin. "Night Work as a Labor Market Phenomenon: Southern Textiles in the Interwar Period." *Explorations in Economic History*, October 1983, 20(4), pp. 331–50.
- Winston, Gordon C. *The timing of economic activities*. New York: Cambridge University Press, 1982.
- Wright, Carroll. *The working girls of Boston*. Boston, MA: Wright and Potter, 1889.

Assortative Mating by Schooling and the Work Behavior of Wives and Husbands

By JOHN PENCAVEL *

In economics, there is an extensive research literature on the market work decisions of husbands and wives, and there is a smaller literature on the marital choices that match husbands and wives. Rarely have these two classes of research been joined. This paper examines the consequences of marital choices for market work behavior. This paper concentrates on one of the dimensions on which men and women are paired in marriage, their schooling.

I. Changes in Assortative Mating by Schooling

It is well known that the years of schooling of husbands and those of wives are positively correlated. Less well known is the fact that this sort of associative mating has changed over the past 50 years. The changes are summarized in Table 1, based on data taken from the "1 in 100 samples" of the 1940, 1960, and 1990 Censuses of Population. This table divides the schooling levels of husbands and wives into five classes according to the major stages of progress through the school system. The ij th entry of each panel of Table 1 records the percentage of husband-wife couples, where the wife's highest schooling level is the i th and the husband's highest schooling level is the j th. The data in Table 1 are based on sampling white wives aged 25-34 years. I concentrate on people aged 25-34 years because they reflect both the schooling and marriage patterns of the period more closely than older people, although the general patterns in these data are also evident (though to a less marked degree) in all those married aged 25-64 years.

The striking feature of Table 1, of course, is the great growth in schooling attainments of

both husbands and wives: in 1940, only about 12 percent of these wives and 15 percent of these husbands had more than 12 years of schooling; by 1990, these percentages were about 55 percent for both husbands and wives. In any Census year, the schooling distribution of wives tends to show less variation than that of husbands: a larger fraction of husbands occupy both the lowest and the highest schooling categories, though the husband-wife difference narrows over time and is least in 1990.

Indicators of the extent of assortative mating in schooling may be derived from the data in Table 1. The simplest and narrowest indicator is the odds of the wife and husband having the same schooling level. This is formed in any year by calculating the ratio of the sum of the entries along the main diagonal to the sum of the entries in all the off-diagonal elements. According to this measure, schooling homogamy (i.e., the correlation between the wife's and husband's schooling) was greatest in 1940 when it was 1.08. Then it fell considerably from 1940 to 1960 when the odds of being married to someone with the same schooling level was 0.78, and it has risen since so that by 1990 the odds of being married to someone with the same schooling was 1.03. Another indicator of assortative mating computes the odds of being married to someone whose schooling differs by at most one level. The movements over time in this indicator are similar: these odds are 5.13 in 1940, 4.68 in 1960, and 8.62 in 1990.

The inference that schooling homogamy has increased since 1960 is subject to the objection that the measures do not effectively hold constant changes in the marginal distributions of schooling. Indeed, changes in homogamy are intimately connected to the increases in schooling levels: if the schooling distribution of husbands and that of wives become less disperse, schooling homogamy in marriage will tend to increase. However, even if one conditions on one spouse's schooling level, sim-

* Department of Economics, Stanford University, CA 94305. I am grateful to David Mancuso for his considerable research assistance. The research reported here was supported by National Science Foundation Grant number SBR-9404482.

TABLE 1—ASSORTIVE MATING ON YEARS OF SCHOOLING

Wife's schooling (years)	Husband's schooling (years)					
	<9	9-11	12	13-15	≥16	All
1940:						
<9	30.7	5.5	2.3	0.5	0.3	39.3
9-11	9.9	8.4	3.6	1.1	0.6	23.6
12	5.7	5.5	8.6	2.7	2.3	24.8
13-15	1.1	1.2	1.8	1.9	2.1	8.1
≥16	0.3	0.2	0.6	0.7	2.3	4.1
All	47.7	20.8	16.9	6.9	7.6	100
1960:						
<9	8.7	3.1	2.0	0.4	0.1	14.3
9-11	6.1	8.0	6.2	1.5	0.6	22.4
12	5.4	9.1	19.1	5.8	4.1	43.5
13-15	0.5	0.9	2.7	2.8	4.6	11.5
≥16	0.1	0.2	0.8	0.9	5.3	7.3
All	20.8	21.3	30.8	11.4	14.7	100
1990:						
<9	0.9	0.5	0.4	0.2	0.0	2.0
9-11	0.7	3.0	3.2	1.5	0.3	8.7
12	0.7	4.0	17.9	9.0	2.9	34.5
13-15	0.3	1.8	8.6	14.5	7.7	32.9
≥16	0.1	0.3	2.3	5.1	14.5	22.3
All	2.7	9.6	32.4	30.3	25.4	100

ilar inferences are derived. For instance, consider all husbands with 16 or more years of schooling and ask, "Among these husbands, what are the odds of their wives having approximately the same schooling (meaning schooling of 13 or more years)?" The answer is that these odds were 2.02 in 1960 and 6.94 in 1990. Or consider all wives with 12 years of schooling and ask, "Among these wives, what are the odds of their husbands having approximately the same schooling (meaning schooling of 9-15 years)?" The answer is that these odds were 3.58 in 1960 and 8.58 in 1990.

Of course, intertwined with changes in the schooling of marriage partners are variations in the length of schooling and the age of marriage. From the 1930's to the 1970's, age at first marriage declined at the same time as schooling levels were rising, so people were more likely to meet their future spouses while still at school. This would make for more schooling homogamy although homogamy tended to fall between 1940 and 1960 and started to increase only after 1960. From the 1970's, the age of first marriage has risen along with time spent at school.

Throughout this period, employment-population ratios of married women have been rising. A man and a woman marrying in the 1980's would have a very much higher expectation that the woman would work for pay than they would in the 1930's. The labor-market returns of women with different levels of schooling introduce a factor in the woman's and man's choice of mate that was of much less importance in the 1930's. And if the wage premia of well-educated men and women rise, as they appear to have done over the past 20 years or so, this changes the monetary implications of a match involving a well-educated man or woman: the forgone consumption from not marrying a well-educated person has risen over the last 15 years.

Therefore, since 1960, young husbands and wives have become more similar in their schooling backgrounds. This conclusion is also reached by Robert Mare (1991) and by David Mancuso (1997) after fitting more sophisticated models designed to discriminate between changes in homogamy and changes in the schooling levels of the entire population.¹ Do these marriage patterns have implications for work behavior? Most basic to theories of systematic choice of marriage partner and of subsequent joint household decision-making is the notion that the behavior of one spouse is related to the characteristics of the other. In the present context, this entails examining the association between the work behavior of one partner and his or her spouse's schooling level, a relationship that has been the subject of little research.

II. Work Hours and Schooling

The research reported in this section is directed toward determining whether the market

¹ Mare (1991) used slightly different schooling categories and different samples, but the general conclusion he reached regarding the movements in assortative mating are similar to those here. He was writing before the 1990 Census became available, and so he made use of three Current Population Surveys in 1985-1987. From the Current Population Surveys, he inferred that, although schooling homogamy increased from 1960 to 1970 and 1980, it appeared not to increase after 1980. The 1990 Census data I have used suggest that homogamy did further increase between 1980 and 1990.

work hours of married men and women are correlated with their spouses' schooling levels. In so doing, we control for other variables, especially the wages of husbands and wives. We examine husbands and wives drawn from the "1 in 100 sample" of the 1990 Census, limiting the sample to couples in which both the husband and wife work for pay. Two pairs of least-squares regression equations are estimated, one pair where the dependent variable is the logarithm of the weekly work hours of the husband and the other pair where the dependent variable is the logarithm of the weekly work hours of the wife. I consider first all husbands and wives. Columns (i) and (iii) of Table 2 report the estimated coefficients on the husband's and wife's schooling dummy variables.

Consider first the hours worked of husbands. Once wages and other variables are held constant, the relationship reported in column (i) between the husband's work hours and his own schooling levels and his wife's schooling levels is weak. A husband with a college education works about 1-percent more hours than his high-school-dropout equivalent while, if he is married to a college-educated woman, he works another 1-percent more than if he were married to a high-school dropout. The own-schooling and cross-schooling correlations with the husband's work hours are of little consequence.

This is not the appropriate inference from the examination of the work behavior of the wife in column (iii) of Table 2. Once her wages and her husband's wages are held constant, a wife's schooling and her work hours are negatively correlated: college-educated wives work 6–8-percent fewer hours than high-school dropouts.² Moreover, women married to well-educated men also tend to work fewer hours: consider two women both with same schooling and whose wages, non-wage income, and husbands' wages are the same; the woman married to a college-educated man works 4-percent fewer market hours than the woman married to a high-

TABLE 2—WORK HOURS AND SCHOOLING OF HUSBANDS AND WIVES

	Dependent variable = logarithm of hours of husband		Dependent variable = logarithm of hours of wife	
	(i) All husbands	(ii) Husbands with children aged <6 years	(iii) All wives	(iv) Wives with children aged <6 years
Schooling (years)				
Husband:				
<12	(reference)	(reference)	(reference)	(reference)
12	0.007 (0.002)	0.014 (0.005)	–0.011 (0.003)	–0.017 (0.009)
13–15	0.015 (0.003)	0.031 (0.005)	–0.026 (0.004)	–0.050 (0.009)
16	0.008 (0.003)	0.020 (0.006)	–0.042 (0.004)	–0.084 (0.011)
>16	0.011 (0.003)	0.022 (0.008)	–0.040 (0.005)	–0.096 (0.013)
Wife:				
<12	(reference)	(reference)	(reference)	(reference)
12	0.002 (0.003)	0.023 (0.006)	–0.022 (0.004)	–0.051 (0.010)
13–15	0.012 (0.003)	0.030 (0.006)	–0.048 (0.004)	–0.095 (0.011)
16	0.011 (0.003)	0.031 (0.007)	–0.079 (0.005)	–0.155 (0.012)
>16	0.013 (0.004)	0.036 (0.008)	–0.056 (0.005)	–0.105 (0.014)

Notes: The numbers reported are least-squares estimated regression coefficients on schooling dummy variables with their associated estimated standard errors in parentheses. The dependent variable is the logarithm of weekly hours of market work. The sample in columns (i) and (iii) consists of 156,715 white husbands and wives, both of whom work for pay, aged between 18 and 64 years. In columns (ii) and (iv), the sample size is 35,702. The samples exclude couples where the husband or the wife has any self-employment income and couples where either worker's estimated hourly wage is less than one dollar or whose earnings are top-coded. Many other variables are included in the regression equation in addition to the schooling dummies: dummy variables for each decade in the husband's and wife's weekly wage distribution; dummy variables for the household's nonwage income (the sum of rent, dividends, and interest); dummy variables for the age of the husband and wife, their school attendance, the presence of young children or no children [excluded from the specification in columns (ii) and (iv)], their location within a central city, and their region of the country. The equations whose estimates are reported in columns (i) and (iii) do not include explicit adjustments for sample selection; when equations were fitted with such adjustments, the point estimates were virtually identical to those reported.

school dropout. At 40 hours per week, 4-percent fewer hours corresponds to a work week of about 38.4 hours.³

² This is not a new finding, although it is not usually given much attention. As an example, see John Cogan (1980) for a previous study with this same result.

³ Of course, if women and men are of the same schooling levels, the effects in Table 2 are additive. For example,

Why should women work less in the marketplace when their spouses are well educated? One explanation derives from the recognition that schooling yields nonmarket as well as market benefits. The contribution of schooling to market productivity is reflected in the higher wage that well-educated people command. But once the effects of schooling on market productivity (as reflected in wages) are held constant, greater schooling indicates greater nonmarket productivity. This higher nonmarket productivity is suggested by the better health that well-educated people exhibit and by the advantages conferred on the children of better-educated parents. Indeed, columns (ii) and (iv) suggest that these cross-schooling effects are stronger among couples with young children.

Column (ii) reports the estimates of the schooling coefficients when the equation is fitted to the sample of husbands, while column (iv) reports the estimates when fitted to the sample of wives who have any children less than six years of age. According to column (iv) these wives work 10-percent fewer hours if married to men with more than 16 years of schooling than if they were married to high-school dropouts. And according to the estimates reported in column (ii) these college-educated husbands substitute some of their own time for their wives' time in the market. For these couples with young children, when both the wife and the husband have a college education, the wife's market work hours are over 20-percent lower than the corresponding hours for those couples where both

the husband and wife did not graduate from high school.⁴

Changes in marriage patterns and the increasing tendency for husbands and wives to share a common schooling background have consequences for work behavior. The issue deserves further study to determine whether the particular pairing of husbands and wives by schooling has effects on market work that augment or offset the relationships measured here.⁵ More generally, the consequences of marital selection for household behavior merits much more extensive investigation.

REFERENCES

- Cogan, John. "Married Women's Labor Supply: A Comparison of Alternative Estimation Procedures," in James P. Smith, ed., *Female labor supply: Theory and estimation*. Princeton, NJ: Princeton University Press, 1980, pp. 90-118.
- Mancuso, David. "Some Implications of Marriage and Assortative Mating by Schooling for the Earnings of Men." Ph.D. dissertation, Stanford University, 1997.
- Mare, Robert D. "Five Decades of Educational Assortative Mating." *American Sociological Review*, February 1991, 56(1), pp. 15-32.
- Pencavel, John. "Schooling and Work of Husbands and Wives." Unpublished manuscript, Stanford University, October 1997.

⁴ Results for other groups in the population are contained in Pencavel (1997).

⁵ In other words, according to the specification reported in Table 2, the effects on the wife's work hours are additive in the husband's schooling and the wife's schooling. In fact, the effect on the wife's work hours of pairing with a college-educated husband may differ between a college-educated wife and a high-school-educated wife.

a college-educated woman married to a college-educated man works about 12-percent fewer market hours than a woman who dropped out of high school and who married a man who was also a dropout.

The Unequal Work Day: A Long-Term View

By DORA L. COSTA *

The length of the work day has fallen sharply over the last century. The typical worker in the 1880's labored ten hours a day whereas his 1940's counterpart worked an eight-hour day. Time-diary studies suggest that the typical worker today works less than eight hours a day (John Robinson and Geoffrey Godbey, 1997 p. 95). The primary beneficiaries of the relatively small declines in the length of the work day since mid-century have been lower-paid workers. Robinson and Godbey (1997 p. 217) note that Americans with a college education work longer hours than Americans with less formal education, and to a lesser extent, those with larger incomes or in professional occupations work the longest hours. Mary T. Coleman and John Pencavel (1993) find that increases in weekly hours of work for the college-educated and declines for those with a high-school education or less have been ongoing since 1940.

Less is known about the distribution of hours worked prior to 1940, the year of the first Census to contain a question on weekly hours worked. Indirect evidence that the distribution of work hours narrowed is available from national consumer expenditure surveys dating back as far as 1888. These show that the difference in recreational expenditures, and hence leisure hours, by social class narrowed sharply prior to 1940 (Costa, 1997), implying that inequality of living standards fell. In contrast, the existing data on trends in wage inequality prior to 1940 (although sometimes contradictory) suggest that wage inequality declined only slightly from the end of the 19th century to 1940 and never fell below

today's levels (Claudia Goldin and Robert A. Margo, 1992). If the lowest-paid workers worked the longest hours in the past, whereas today it is the most highly paid who work the longest hours, then wage or wealth data may underestimate long-run improvements in the welfare of the lowest-income workers and may present a skewed picture of recent trends in the inequality of living standards.

I. Who Worked the Longest Day?

The data that reveal the distribution of hours worked in the past come from the numerous surveys of the personal, occupational, and economic circumstances of nonfarm wage-earners published by state bureaus of labor statistics in the last quarter of the 19th century.¹ Although the surveys are predominately of upper-working-class men, when the data sets that provide information on men's daily hours of work, their wages, and their age are pooled to yield a sample of over 11,000 men aged 25–64 there is enough variation in the data to re-weight by broad occupational or industry category.

The questions that were asked about hours of work varied slightly by state, but all referred to usual hours of work per day. The mean length of the work day in the pooled data set is about 10 hours (even when the data are re-weighted to be representative of either the 1900 or the 1910 industrial distribution), an estimate similar to that obtained from other sources, such as the 1880 manufacturing census (Jeremy Atack and Fred Bateman, 1992) and the Reports of the Commissioner of Labor

* Department of Economics, Massachusetts Institute of Technology, E52-274C, 50 Memorial Drive, Cambridge, MA 02142-1347, and National Bureau of Economic Research. This research was funded by NIH grant AG12658. I have benefited from the comments of Matthew Kahn, Peter Temin, and workshop participants at Harvard University, and at Duke University, North Carolina State University and the University of North Carolina.

¹ Many of these surveys were collected by Susan Carter, Roger Ransom, Richard Sutch, and Hongcheng Zhao and are available on the World Wide Web from <http://www.eh.net/Databases/Labor/>. The surveys used in this study are California in 1892; Kansas in 1895, 1896, 1897, and 1899; Maine in 1890; Michigan stone workers in 1888; Michigan railway workers in 1893; and Wisconsin in 1895.

(Robert Whaples, 1990 p. 33). A comparable measure of usual daily hours of work is provided by the 1991 May supplement to the Current Population Survey and can be inferred from usual hours per week divided by usual days per week as given in the 1973 May supplement. The mean number of daily hours of work in both years was 8.6, and although the coefficient of variation increased somewhat from the 1890's to 1973, the distribution between the 90th and 10th percentiles has become more compressed because the majority of workers now work an eight-hour day.

Table 1 gives average hours worked per day by deciles of the average hourly wage, both for men paid by the hour and for all men aged 25-64 in the 1890's, 1973, and 1991. Note that in the 1890's hours worked decrease sharply for men in higher deciles. By 1973 the decrease in hours was no longer as pronounced. By 1991 daily hours increase with the wage decile and then level off, rising only slightly at a higher wage decile. Even within wage deciles and within occupational and industry categories the lower-paid workers worked the longest day in the 1890's, whereas the higher-paid workers worked the longest day in 1991. The pattern persists when one controls for age, marital status, number of dependents, and state and year fixed effects as well.

Although the required micro data do not exist to ascertain exactly when the compression between the 1890's and 1973 occurred, the trend in the mean length of the work day suggests that most of the change occurred by the mid 1920's. Because the decline in hours worked between the 1890's and 1973 was largest among men earning the lowest wages, most changes in the mean length of the work day must have come from disproportionate declines in the hours of men in the lowest deciles of the wage distribution, and by 1926, if not earlier, the length of the work day for manufacturing workers was around 8.3 hours (Paul H. Douglas, 1930).

II. Explanations

Various factors might account for the change in the distribution of daily hours worked by different wage deciles. The number

TABLE 1—DISTRIBUTION OF USUAL-LENGTH WORK DAY BY HOURLY WAGE DECILES, MEN AGED 25-64, 1890's, 1973, AND 1991

Wage decile	A. All workers		
	1890's	1973	1991
<10 (bottom)	10.99	8.83	8.05
10-20	10.46	8.47	8.47
20-30	10.50	8.54	8.53
30-40	10.63	8.38	8.61
40-50	10.31	8.34	8.59
50-60	9.99	8.33	8.61
60-70	10.29	8.33	8.47
70-80	10.07	8.32	8.66
80-90	9.64	8.26	8.64
≥90 (top)	8.95	8.22	8.72
90th/10th	0.81	0.93	1.08
90th/50th	0.90	0.99	1.01
50th/10th	0.94	0.94	1.07

Wage decile	B. Paid by hour		
	1890's	1973	1991
<10	11.14	8.17	7.64
1-20	10.08	8.23	8.14
20-30	9.62	8.23	8.24
30-40	9.62	8.16	8.30
40-50	9.62	8.12	8.38
50-60	9.33	8.15	8.48
60-70	9.42	8.16	8.26
70-80	8.67	8.20	8.47
80-90	8.50	8.15	8.40
≥90	8.88	8.01	8.51
90th/10th	0.80	0.98	1.11
90th/50th	0.95	0.98	1.00
50th/10th	0.86	0.99	1.10

Notes: The 1890's data are weighted to have the same distribution of occupational categories as the population in 1900. For workers paid by the hour, the hourly wage is the reported hourly wage. For 1973 and 1991, the hourly wage for all workers was estimated from information on weekly earnings. For the 1890's, this rate had to be estimated for workers paid by the week, month, or year by assuming a standard work year, month, or week and subtracting the number of days lost due to ill health, unemployment, or other factors. Although this imputation procedure might introduce systematic bias, examining workers paid by the hour provides some indication of the direction and magnitude of the bias.

of daily hours supplied by men in the lowest wage decile may have fallen relative to the number of hours supplied by men in the top decile. Technological change such as electrification which allows firms to use different

shifts of workers may have decreased firms' demand for daily hours from each individual worker, but disproportionately so for hours of work of lower-skilled and hence lower-paid workers. Hours legislation may have lowered the hours worked by men in the lowest wage deciles. Lastly, the distribution of daily hours may be a poor indicator of total or yearly hours.

A. Demand and Supply

If some of the occupations or industries that experienced large hours declines were the occupations or industries that employed many low-decile workers, then hours of workers in the lowest deciles may have fallen simply because they were overrepresented in the occupations or industries that experienced declines in hours. In the 1890's professionals, craft workers, and laborers worked a much shorter day compared to managers, service, and sales workers. Classifying men by industry shows that the longest hours were worked in trade and personal service and the shortest in mining and construction. By 1991, managers and sales workers still worked the longest day, but service and clerical workers worked the shortest day. The longest working days were in mining, transportation, communication, utilities, and trade, and the shortest were in entertainment and personal service. Because hours worked are not known for some industries in the 1890's, I only analyze demand shifts due to broad interoccupational hours changes.

The horizontal shift in demand for daily hours of work from an individual in wage-decile i due to changes in the interoccupational mix of daily hours at fixed wages is

$$\Delta h_i = \sum_j \alpha_{ij} a_{ij} \Delta H_j$$

where H_j is the average number of daily hours worked in occupation or industry j , a_{ij} is the ratio of daily hours worked in wage decile i to average occupation or industry hours (H_{ij}/H_j), α_{ij} is the fraction of workers in wage decile i in occupation or industry j , and α_{ij} and a_{ij} are evaluated at the base year.

I determine changes in the supply of daily hours worked of men in a given wage decile by explicitly estimating labor-supply equa-

tions for each period and then using the estimated regressions to predict daily hours of work within each wage decile. The equations that I estimate are

$$(1) \quad h_i = \beta_0 + \beta_w w_i + \mathbf{x}_i' \boldsymbol{\beta}$$

for the 1890's and

$$(2) \quad h_i = \beta_0 + \beta_w \ln(w_i) + \mathbf{x}_i' \boldsymbol{\beta}$$

for 1991 and 1973, where h is hours worked, w is the hourly wage, and \mathbf{x} is a vector of demographic characteristics, such as age and number of dependents. Endogeneity between the wage and hours presents potential problems. Because individuals may influence their own wage through investment in human capital, the wage is likely to be correlated with the stochastic error term due to unobserved tastes and abilities that help determine the wage and that determine current labor supply. I therefore use industry dummies as instruments under the assumption that hours demanded from each worker depend upon the industry (perhaps because of technological factors) and hours supplied by each worker do not.

Table 2 presents estimates of wage elasticities for the 1890's, 1973, and 1991. Note that the supply curve of daily hours in the 1890's was very backward-bending, consistent with other estimates for the period (e.g., Whaples, 1990) and with contemporary observations. The estimates are more strongly negative than those obtained from ordinary least squares. I obtain negative labor-supply elasticities conditioning on broad occupational group as well.

In contrast to the estimates for the 1890's, elasticities estimated for 1973 and 1991, although negative, are fairly small. For workers paid by the hour they are positive in 1991. Of course, the negative labor-supply elasticities estimated for 1973 and 1991 may well be spurious. Estimates derived from modern panel data sets suggest that, upon instrumenting, changes in hours worked are positively related to increases in wages, even when ordinary least-squares analysis indicates that the relationship is negative (e.g., Shelly Lundberg, 1985). Regardless of whether recent labor-supply elasticities are positive or slightly negative but small, the comparison with past

TABLE 2—ELASTICITY (INSTRUMENTAL-VARIABLES ESTIMATES) OF DAILY HOURS WORKED WITH RESPECT TO THE HOURLY WAGE, MEN AGED 25–64, 1890's, 1973, AND 1991

Instruments	Wage elasticity	
	All workers	Hourly workers
Industry dummies		
1890's	–0.304 (0.023)	–0.536 (0.126)
1973	–0.087 (0.013)	–0.023 (0.011)
1991	–0.017 (0.016)	0.104 (0.019)

Notes: Standard errors are in parentheses. Control variables for the 1890's data are age, age-squared, dummies for foreign birth, home-ownership, whether the worker has any dependents, and fixed effects indicating which state bureau of labor statistics report the data came from. Control variables for 1973 and 1991 are age, age-squared, dummies for nonwhite and married, and state fixed effects; for 1991 only, the number of children under age 18 is an additional control variable. All elasticities are estimated at the variable means.

labor-supply elasticities suggests that at least between the 1890's and 1973 the labor-supply curve has become less backward bending.

Table 3 summarizes changes in supply and demand shifts for daily hours of men in the top wage decile relative to those of men in the bottom wage decile. Assuming that total demand shifts were proportional to the estimated partial demand shifts, Table 3 suggests that changes in labor supply dominated changes in labor demand.

B. Other Factors

Although hours legislation may explain much of the current concentration on eight hours of work, it cannot explain most of the compression in the distribution of the length of the work day. Recall that I argued that most of the compression probably occurred by the mid-1920's. But, prior to the 1930's state legislation restricting maximum hours of work applied only to women and to relatively few men in dangerous industries.

Workers' trading-off a long work day for a short work week or year is also an unlikely explanation. In the 1890's workers who re-

TABLE 3—ANNUAL DEMAND AND SUPPLY SHIFTS FOR DAILY HOURS OF WORKERS IN THE TOP DECILE RELATIVE TO DAILY HOURS OF WORKERS IN THE BOTTOM DECILE

Comparison	Relative shift	
	Demand	Supply
All Workers		
1973 vs. 1890	0.04	0.30
1991 vs. 1973	0.08	1.26
Paid by hour		
1973 vs. 1890		0.39
1991 vs. 1973		1.33

Notes: See the text for details. Supply shifts were estimated using predicted average daily hours from the industry instrumental-variables specifications.

ported that Sunday work was required were more likely to work a longer day, as were those who reported either no reduction or an increase in Saturday hours. In 1991, men who worked a longer usual day reported working longer usual weekly hours and more days per week. That declines in seasonality have led to less substitution by workers of a longer work day for downtime can be ruled out as well. The number of days lost by the individual worker in the past year has a negative, but negligible, effect on his usual hours of work. Controlling for observable characteristics such as the wage and demographic information, workers in occupations where mean unemployment was three months in the year labored almost two hours less per day than workers in occupations with zero mean unemployment.

III. Implications

The distribution of hours worked has implications for trends in income inequality. Table 4 shows that, between 1973 and 1991, $[(1.22 - 1.16)/(1.39 - 1.16)] \times 100 = 26$ percent of the earnings inequality between the 90th and the 10th wage deciles could be attributed to differences in hours worked. Table 4 also shows that had the 1991 pattern of hours worked prevailed in the 1890's (but with the number of days worked per week remaining unchanged) weekly earnings inequality would have been much greater in the 1890's than it actually was. In the past, an inequalitarian

TABLE 4—WEEKLY EARNINGS INEQUALITY, 1890's, 1973, AND 1991

Deciles	Difference in log weekly earnings				
	Actual			At 1991 Hours	
	1890's	1973	1991	1890's	1973
90th vs. 10th	1.13	1.16	1.39	1.36	1.22
90th vs. 50th	0.57	0.56	0.65	0.68	0.59
50th vs. 10th	0.56	0.60	0.73	0.67	0.63

Notes: The 1890's data are weighted by the distribution of occupational groups in the 1900 population; nonetheless, because the 1890's data are not a random sample of the population, wage inequality in the 1890's may be underestimated. Weekly earnings in the 1890's were estimated assuming a regular work year of 307 days minus days lost due to unemployment, sickness, or other causes.

distribution of work tended to equalize income, whereas today it magnifies earnings disparities.

Changes in the structure of daily work hours could largely be accounted for by declines in the relative number of daily hours workers were willing to supply. Compared to the 1890's increases in the hourly wage no longer have a large, negative impact on hours worked. In fact, workers are now slightly more willing to increase their hours as their wages rise. Several factors could explain this change. Because the work day is now so much shorter, workers are no longer as time-poor. Because their incomes are now higher, the income effect of a wage increase could be smaller. The availability of new consumer goods might have increased demand for goods relative to leisure. The cost of recreation may have fallen disproportionately for lower-income workers. Alternatively, workers may now prefer to take their leisure at older ages and work longer hours during their prime, perhaps because it is now easier for them to save for their retirement since Social Security and private pension plans provide strong financial in-

centives to take leisure at older ages, or because leisure at older ages is now much more fun. Regardless of the reason for the change in the distribution of work hours the results of this paper imply that, although the rich and the poor will always differ in terms of income, income differences no longer mean that the poor have less time for fun.

REFERENCES

- Atack, Jeremy and Bateman, Fred. "How Long Was the Workday in 1880?" *Journal of Economic History*, March 1992, 52(1), pp. 129–60.
- Coleman, Mary T. and Pencavel, John. "Changes in Work Hours of Male Employees, 1940–1988." *Industrial and Labor Relations Review*, January 1993, 46(2), pp. 262–83.
- Costa, Dora L. "Less of a Luxury: The Rise of Recreation Since 1888." National Bureau of Economic Research (Cambridge, MA) Working Paper No. 6054, June 1997.
- Douglas, Paul H. *Real wages in the United States, 1890–1926*. Boston, MA: Houghton Mifflin, 1930.
- Goldin, Claudia and Margo, Robert A. "The Great Compression: The Wage Structure in the United States at Mid-Century." *Quarterly Journal of Economics*, February 1992, 107(1), pp. 1–34.
- Lundberg, Shelly. "The Tied Wage–Hours Offers and the Endogeneity of Wages." *Review of Economics and Statistics*, August 1985, 67(3), pp. 405–10.
- Robinson, John and Godbey, Geoffrey. *Time for life: The surprising ways Americans use their time*. University Park, PA: Pennsylvania State University Press, 1997.
- Whaples, Robert. "The Shortening of America's Work Week: An Economic and Historical Analysis of its Context, Causes, and Consequences." Ph.D. dissertation, University of Pennsylvania, 1990.

WHAT IS POVERTY AND WHO ARE THE POOR? REDEFINITION FOR THE UNITED STATES IN THE 1990's[†]

Absolute versus Relative Poverty

By JAMES E. FOSTER *

Should poverty be measured using an "absolute" or a "relative" approach? This age-old question in poverty measurement is once again on the agenda, due to the ambitious proposals of Patricia Ruggles (1990) and the National Research Council of the National Academy of Sciences (Constance Citro and Robert Michael, 1995) to alter the way U.S. poverty is measured. Their wide-ranging suggestions include a new "hybrid" approach to setting the poverty threshold that, unlike the current absolute method, is sensitive to changes in the general living standard, but less sensitive than a purely relative approach. The proposals also recommend using aggregate indexes of poverty beyond the usual "headcounts," such as well-known "gap" measures and indicators of the distribution of resources among the poor. Important relative notions of poverty enter at this "aggregation" step as well. The effects of the various recommendations on the trend and cross-sectional profiles of poverty are actively being explored (see e.g., David Betson and Jennifer Warlick, 1997; Thesia Garner et al., 1997; David Johnson et al., 1997). At the same time it may prove useful to consider some of the conceptual measurement issues arising from the proposals. This is the direction taken in the present study.

This paper evaluates the multiple notions of relative and absolute poverty that arise in

choosing poverty lines and in aggregating the data into an overall index of poverty. A general taxonomy is presented, and the question of robust comparisons is addressed within this general framework. Special attention is paid to distinguishing between (i) the general concept underlying the poverty line and (ii) the particular cutoff chosen. The paper concludes with a discussion of "hybrid" poverty lines and the associated parameter that is likely to play a key role in future discussions: the income elasticity of the poverty line.

I. Elements

Poverty measurement is based on a comparison of resources to needs. A person or family is identified as poor if its resources fall short of the poverty threshold. The data on families are then aggregated to obtain an overall view of poverty.

There are many ways of defining resources, constructing thresholds, and aggregating the resulting data (see e.g., Ruggles, 1990; Martin Ravallion, 1994; Citro and Michael, 1995). Virtually all partition the population into groups of families (or resource-sharing units) with similar characteristics, and I follow this approach here. Let Θ denote the raw data, containing information on resources received by families, their demographic and other characteristics, and perhaps other data (e.g., consumption distributions) needed to construct poverty thresholds. Let m be the number of distinct groups, with $n^k = n^k(\Theta)$ being the number of families in group k . Once a specific definition of family resources has been fixed, this yields a distribution of resources among the families in group k , denoted by the n^k -dimensional vector $\mathbf{x}^k = \mathbf{x}^k(\Theta)$. The poverty threshold for families in group k is denoted by the number $z^k = z^k(\Theta)$; a family is identified

[†] Discussants: David S. Johnson, U.S. Bureau of Labor Statistics; Patricia Ruggles, U.S. Department of Health and Human Services; Barbara Wolfe, University of Wisconsin; Christopher Jencks, Harvard University.

* Department of Economics, Vanderbilt University, Nashville, TN 37235. I thank the discussant, David S. Johnson, for his insightful commentary. Financial support from the John D. and Catherine T. MacArthur Foundation through the Network on Inequality and Poverty in Broader Perspective is gratefully acknowledged.

as poor if its resource level falls below z^k . Exactly how z^k is to be set (i.e., the "identification step" of Amartya Sen, 1976) is a key part of the present discussion.

As for the "aggregation step," most U.S. studies report poverty levels for the demographic groups and then aggregate to obtain an overall level of poverty. Thus, they implicitly take the poverty index to be "decomposable" across the groups (on which more will be said presently). With overall poverty a weighted sum of group poverties, the aggregation question reduces to a choice of the poverty index $P(\mathbf{x}; z)$ to apply to a typical group distribution \mathbf{x} and poverty line z . The most common index is the head-count ratio $H(\mathbf{x}; z) = q/n$ where q is the number of poor families in \mathbf{x} given z , and n is the number of families in \mathbf{x} . This index provides important information on poverty (namely, the frequency of poverty among the population) but ignores other relevant information on the depth and distribution of poverty. Another important kind of "partial index" is based on the sum of the income gaps $(z - x_i)$ of poor families. These "gap indexes" add a second dimension of "depth" to poverty evaluations. A third dimension is provided by indexes of inequality among the poor.

While each partial index conveys useful information about some aspect of poverty (assuming, of course, that the poverty threshold itself is meaningful), one must be careful in using its unidimensional prescriptions as a guide to policy. For this and other reasons, it has been argued (see e.g., Foster and Sen, 1997) that an index combining all three dimensions is more coherent in this role. Such "distribution-sensitive" indexes have been used to great advantage in international comparisons and development studies (see e.g., Ravallion, 1994).

II. Absolutes and Relativities

There are several ways in which relative and absolute considerations enter into poverty measurement. I offer a simple taxonomy including the *threshold* and *equivalence-scale* choices in the identification step, and the treatment of *population*, *scale*, and *individual deprivation* in the aggregation step.

A. Threshold

The first and perhaps most important sense in which poverty measurement is absolute or relative concerns the setting of the poverty standard. An *absolute* poverty line is a fixed (group-specific) cutoff level z_a that is applied across all potential resource distributions. In comparisons over time, for example, the standard is unchanged even in the face of economic growth (although provisions are made for changes in price levels);¹ similarly, in comparisons across countries, fixed-threshold comparisons require an appropriate exchange rate. If the absolute standard is truly independent of the current data, though, how can one be sure that the standard chosen is an appropriate one? The poverty line is typically calibrated in some initial period using, say, food-budget studies, and it is then carried forth from year to year, irrespective of whether the same procedure applied to current data would yield the same result. In a growing economy, the gap between the hypothetical recalibrated level and the historical standard may well be quite large. Such is the case with the current U.S. poverty standard, and this is one of the criticisms that have been leveled against it (see Citro and Michael, 1995 pp. 2–3).

In contrast, a relative approach uses current data to generate the current poverty threshold. A *relative* poverty line begins with some notion of a *standard of living* $r(\mathbf{x})$ for the distribution \mathbf{x} , such as the mean, median, or some other quantile, and defines the cutoff as some percentage α of this standard. The result is a poverty threshold $z_r = \alpha r(\mathbf{x})$ that varies one-for-one with the standard of living, in that a 1-percent increase in r is matched by a 1-percent increase in z_r . Examples include the "50 percent of the median" relative poverty line proposed by Victor Fuchs (1969) and the "50 percent of the mean" threshold employed by

¹ There is a significant issue of whether resource should be expressed in real terms and, if so, which cost of living index to use. This issue is ignored here for simplicity, but it is clearly another potentially important source of "relativity" in the measurement of poverty.

Michael O'Higgins and Stephen Jenkins (1990).²

Using a relative line does not amount to measuring inequality (although theorem 6 in Foster and Anthony F. Shorrocks [1988a] provides one important link) nor does it imply that poverty is by definition "always with us" (see Anthony Atkinson, 1975 p. 189). And while many studies regard absolute lines as being especially low and relative lines as being high, this is not necessarily the case. If living standards are rising and thresholds are pegged at $z_a = z_r$ in some initial period, then $z_a < z_r$ for all subsequent periods, but $z_a > z_r$ for all previous periods, as emphasized by Citro and Michael (1995 p. 132). In any isolated period, it is not possible to tell whether a given threshold z is relative or absolute, nor is the distinction particularly important, since the same numerical cutoff, however originally derived, must lead to the same level of poverty.

The key distinction between absolute and relative thresholds is not seen in the specific values obtained at a given date, but in how the values change as the distribution changes. Thus, there is an important distinction to be made between the general *concept* underlying the poverty threshold, and the specific *cutoff* selected. For comparisons involving extended periods of time, or very different standards of living, the former is likely to be the more important issue (see also Ruggles, 1990 Ch. 3), while the latter choice (of cutoff) is largely arbitrary (see Fuchs, 1969; Atkinson, 1975, 1987; Foster and Shorrocks, 1988b). This inevitable arbitrariness casts doubt on the meaning of the cardinal poverty levels obtained at specific cutoffs and leads to a consideration of the robustness of results to changes in the cutoff, a topic I will return to below.

B. Equivalence Scale

A second entry point for relativities in poverty measurement is where poverty lines are

adjusted across demographic groups. One approach is to apply repeatedly the procedure for setting poverty lines to each group separately and thereby arrive at m independent thresholds. However, as noted by Ruggles (1990 Ch. 4), this can lead to odd (nonmonotonic) behavior of the poverty line as family size changes. An alternative approach sets the line in one reference group and then derives the remaining thresholds using an "equivalence scale" to account for the differing needs of different-sized families. The typical scale provides the rate at which a dollar for one group translates into dollars for another. So if group 1 is the reference, and s^k is the conversion rate from group 1 to group k , then $z^k = s^k z^1$ becomes the poverty line for group k .

This sort of equivalence scale is *relative* in that the transformation from group to group is multiplicative, and consequently group poverty lines are proportionate to each other. Another possibility raised by Charles Blackorby and David Donaldson (1994) is for variations in family configuration to have an constant *absolute* effect so that, for example, adding another child is seen as an additional fixed (real) cost to the family, independent of the size of the base threshold. Relative equivalence scales preserve the ratios of group poverty lines as the base threshold changes; an absolute equivalence scale preserves the absolute differences. The two forms are indistinguishable for a single observation or if the reference threshold remains unchanged (as with an absolute poverty line).

C. Population

The aggregation stage uses three notions of absolute and relative poverty in constructing poverty indexes. First, a *relative* or per capita poverty index is independent of the population size in the sense that "replicating" the population leaves the poverty value unaffected: for example, $P(\mathbf{x}, \mathbf{x}; z) = P(\mathbf{x}; z)$. Such a measure is based purely on the relative frequencies of incomes in the income distribution. In contrast, an *absolute* index is one whose value rises in proportion to the number of replications: for example, $P(\mathbf{x}, \mathbf{x}; z) = 2P(\mathbf{x}; z)$. The head-count ratio q/n is relative in this sense while the head-count q is absolute. An

² There are important measurement issues in selecting the standard of living. Should it be the mean, the median, or some other representative income? Should it be from the entire population or some reference group? Should it be for all expenditures or a significant subset? (For references, see Citro and Michael [1995].)

absolute index can be converted to a relative index by dividing by n .

D. Scale

A second notion concerns the behavior of an index when the poverty line and incomes are simultaneously altered. A *relative* or scale-invariant index is one that is unchanged when the poverty line and all incomes are multiplied by the same factor. An *absolute* or translation-invariant index is independent of additions of the same constant to the poverty line and all incomes (Blackorby and Donaldson, 1980). Thus, for example, the aggregate poverty gap $\sum_{i=1}^q (z - x_i)$ is an absolute index of this sort, while the normalized poverty gap $\sum_{i=1}^q (z - x_i)/z$ (which measures the poverty gap in poverty line units) is relative. The head-count ratio is both absolute and relative in this sense (and is essentially unique in this respect [see Buhong Zheng, 1994]).

E. Deprivation

Finally, the basic notion of deprivation that underlies a given index may be relative or absolute. If a family's poverty level depends purely upon its own characteristics, its resource level, and its threshold, then the index is based on a notion of *absolute deprivation*. Foster and Shorrocks (1991) relate this to *decomposability* of the index across population subgroups (overall poverty is a weighted sum of subgroup poverties for any partition) and also to a more fundamental notion of *subgroup consistency* (overall poverty is increasing in subgroup poverty levels for any partition). The head-count ratio and the gap indexes are absolute in this sense, as is the index of Foster et al. (1984) which takes $[(z - x_i)/z]^2$ as the i th poor family's deprivation level. In contrast, Sen's (1976) index is founded on the notion of *relative deprivation*, since a family's deprivation level depends crucially on its relative position among poor families and thus incorporates information beyond its own data. A discussion of the two approaches can be found in section A6 of Foster and Sen (1997).

III. Robust Comparisons

The above taxonomy presents several avenues for relative and absolute concepts to enter into poverty evaluations, and many combinations are possible. For example, the current method for evaluating U.S. poverty employs an absolute threshold for each group and a relative or absolute equivalence scale (indeterminate since poverty lines are unchanging) to identify the poor; for the aggregation step it typically uses the head-count ratio, a population relative index that is both absolute and relative with respect to scale, and which is based purely on absolute deprivation. The aggregate gap, which is absolute in all three dimensions, is often used as an alternative index.

Each combination of absolutes and relative concepts has many possible implementations (i.e., specific cutoffs, scales, and indexes) from which to choose. Inevitably, this entails making choices for which there is little guidance (why 50 percent of the median instead of 49 percent?). It is important to note, however, that the decision need not be based on normative or subjective considerations. The selection from the array of possible implementations could be purely arbitrary—made in the interest of getting on with the analysis (on this distinction, see Sen [1980]).

Given the inherent arbitrariness in selecting a specification, it is important to evaluate the robustness of any conclusions obtained. In cases where the numerical poverty levels are important, this may be as simple as testing other reasonable specifications and reporting how the poverty level changes. Betson and Warlick (1997), for example, use 20-percent changes in z to illustrate the cardinal sensitivity of head-counts to the threshold. Alternatively, when rankings of poverty levels are all that matter, one has available a rather large collection of tools to evaluate ordinal robustness (analogous to the well-known Lorenz criterion for inequality analysis), which cover variable thresholds, equivalence scales, and indexes (see e.g., Foster, 1984; Foster and Shorrocks, 1988b; Atkinson, 1987, 1992). Virtually all approaches trace back to notions of stochastic dominance from risk theory (see the general discussion in Foster and Sen [1997]).

Most results of this type are presented in a one-group framework with absolute thresholds; but in fact, the tools have far greater applicability. As an illustration, suppose that the base threshold z^1 and equivalence scale s^k are relative, the index P is based on a notion of absolute deprivation (hence decomposable) but otherwise relative, and the only question is the specific cutoff α to be used in setting the relative poverty line. Suppose that for a specific value of α , say, $\alpha = 50$ percent, the resource distribution (x^1, \dots, x^m) has greater poverty than (y^1, \dots, y^m) . When can one be sure that this will remain true for an entire range of α values, say $(0, \bar{\alpha})$ where $\bar{\alpha} > 50$ percent? Let r be the standard of living underlying the relative poverty line, and let r_x and r_y denote the respective standards in the distributions. Construct a new "equivalent" distribution \tilde{x}^k for demographic group k by dividing family resources by the equivalence scale s^k , and then replicating by family size in k , so that \tilde{x}^k has one equivalent resource level for each person in group k . It is not difficult to show that for P satisfying the above properties, the poverty level of the original distribution (x^1, \dots, x^m) at the group-specific thresholds $z^k = s^k z^1$ is simply

$$P(\tilde{x}^1, \dots, \tilde{x}^m; z^1)$$

or the poverty in the equivalent distribution given group 1's poverty line. If one further normalizes incomes by the standard of living, then the poverty level is given by

$$P(\tilde{x}^1/r_x, \dots, \tilde{x}^m/r_x; \alpha).$$

Consequently, the judgment that (x^1, \dots, x^m) has greater poverty than (y^1, \dots, y^m) is in fact robust in α if

$$\begin{aligned} &P(\tilde{x}^1/r_x, \dots, \tilde{x}^m/r_x; \alpha) \\ &> P(\tilde{y}^1/r_y, \dots, \tilde{y}^m/r_y; \alpha) \end{aligned}$$

for all $\alpha \in (0, \bar{\alpha})$.

This last condition is in a form that allows the application of results in Foster and Shorrocks (1988b) and Atkinson (1987). So, for example, the test for the head-count ratio H checks whether the two distributions of nor-

malized equivalent incomes can be compared using first-degree stochastic dominance over the range $(0, \bar{\alpha})$, while the tests for the normalized gap index and the Foster et al. (1984) index use second- and third-degree stochastic dominance, respectively. Atkinson's (1987) results go beyond these results to consider variations in poverty indexes and indicate, for example, that if there is an unambiguous comparison for H (and hence first-degree stochastic dominance), then virtually any acceptable index P will agree with this conclusion. This illustrates the power of the head-count ratio in this context.

IV. Hybrid Measurement

Many of the categories in my taxonomy allow for an intermediate position to be chosen in place of a pure relative or absolute approach. One particularly interesting example is the "hybrid" poverty threshold that is central to the proposal in Citro and Michael (1995), which is based on what might be termed a "partial" standard of living: r_p is the median expenditure on certain basic goods. The threshold $z = \alpha r_p$ has the same structure as a purely relative cutoff (and in fact the robustness result applies equally well to it). However, median expenditures on basic goods do not rise as fast as, say, median total expenditures, and it is this empirical fact that gives z its hybrid nature.

One could also imagine thresholds that are hybrid by construction, in that they depend directly on an absolute and a relative standard. For example, consider a weighted geometric average of a relative threshold $z_r = \alpha r$ and an absolute threshold z_a , namely, $z = z_r^\rho z_a^{1-\rho}$, where $0 < \rho < 1$. This form of hybrid line has the property that a 1-percent increase in the living standard r always leads to a ρ -percent increase in the poverty line. In other words, ρ is the elasticity of the poverty line with respect to the living standard, or what Gordon Fisher (1995) has termed the *income elasticity of the poverty line*. In general, $\rho = (dz/dr)(r/z)$ has a natural interpretation as a measure of the extent to which a given threshold z is relative, with $\rho = 0$ corresponding to an absolute poverty line and $\rho = 1$ a fully relative one. The possibility of using a hybrid standard changes

the question "absolute or relative?" to "exactly how relative?" with ρ as the relevant decision variable.

In his defense of the relative approach, Fuchs (1969 p. 201) posited that the cutoff "would be recognized as a national value judgment and would be arrived at through the normal political process." One theme of the present paper is the primacy of general concept over specific cutoff; if this is accepted, then the subject of public discourse would more properly be ρ , the income elasticity of the poverty line. The choice of ρ would then answer the normative question: "To what extent should the poor share in economic growth?" An elasticity of 1 appears to be too high to command much political support in the United States. An elasticity of 0 is implicit in the current standard, but given the historical tendency for absolute standards to be periodically revised (Fisher, 1995) and the long-standing explanations why, when the general standard of living rises, resources may need to be higher to achieve the same ends (e.g., Atkinson, 1975; Sen, 1983), this answer may not be tenable in the long run. However, it remains to be seen whether the particular hybrid standard proposed by National Research Council of the National Academy of Sciences, which has a historical income elasticity of $\rho = 0.65$ (Citro and Michael, 1995 p. 143), will garner enough support to displace the current standard.

REFERENCES

- Atkinson, Anthony B. *The economics of inequality*. Oxford: Oxford University Press, 1975.
- . "On the Measurement of Poverty." *Econometrica*, July 1987, 55(4), pp. 749–64.
- . "Measuring Poverty and Differences in Family Composition." *Economica*, February 1992, 59(233), pp. 1–16.
- Betson, David M. and Warlick, Jennifer L. "Alternative Historical Trends in Poverty." Mimeo, University of Notre Dame, 1997.
- Blackorby, Charles and Donaldson, David. "Ethical Indices for the Measurement of Poverty." *Econometrica*, May 1980, 48(4), pp. 1053–60.
- . "Measuring the Cost of Children: A Theoretical Framework," in Richard Blundell, Ian Preston, and Ian Walker, eds., *The measurement of household welfare*. Cambridge: Cambridge University Press, 1994, pp. 51–69.
- Citro, Constance F. and Michael, Robert T. *Measuring poverty: A new approach*. Washington, DC: National Academy Press, 1995.
- Fisher, Gordon M. "Is There Such a Thing as an Absolute Poverty Line Over Time?" Mimeo, U.S. Department of Health and Human Services, Washington, DC, August 1995.
- Foster, James E. "On Economic Poverty," in Robert Basmann and George Rhodes, eds., *Advances in econometrics*, Vol. 3. Greenwich, CT: JAI Press, 1984, pp. 215–51.
- Foster, James E.; Greer, Joel and Thorbecke, Erik. "A Class of Decomposable Poverty Measures." *Econometrica*, May 1984, 52(3), pp. 761–66.
- Foster, James E. and Sen, Amartya. "On Economic Inequality after a Quarter Century," in Amartya Sen, ed., *On economic inequality*. Oxford: Clarendon, 1997, pp. 107–219.
- Foster, James E. and Shorrocks, Anthony F. "Inequality and Poverty Orderings." *European Economic Review*, March 1988a, 32(2–3), pp. 654–61.
- . "Poverty Orderings and Welfare Dominance." *Social Choice and Welfare*, 1988b, 5(2/3), pp. 179–98.
- . "Subgroup Consistent Poverty Measures." *Econometrica*, May 1991, 59(3), pp. 687–709.
- Fuchs, Victor. "Comment on Measuring the Size of the Low-Income Population," in Lee Soltow, ed., *Six papers on the size distribution of wealth and income*. New York: National Bureau of Economic Research, 1969, pp. 198–202.
- Garner, Thesia I.; Paulin, Geoffrey; Shipp, Stephanie; Short, Kathleen and Nelson, Chuck. "Experimental Poverty Measurement for the 1990's." Mimeo, Bureau of Labor Statistics, Washington, DC, 1997.
- Johnson, David; Shipp, Stephanie and Garner, Thesia. "Developing Poverty Thresholds Using Expenditure Data." Unpublished manuscript presented at the Joint Statistical Meetings, Anaheim, CA, August 1997.

- O'Higgins, Michael and Jenkins, Stephen. "Poverty in the EC: Estimates for 1975, 1980, and 1985," in Rudolph Teekens and Bernard M. S. van Praag, eds., *Analysing poverty in the European Community: Policy issues, research options, and data sources*. Luxembourg: Office of Official Publications of the European Communities, 1990, pp. 187-212.
- Ravallion, Martin. *Poverty comparisons*. Chur, Switzerland: Harwood, 1994.
- Ruggles, Patricia. *Drawing the line*. Washington, DC: Urban Institute Press, 1990.
- Sen, Amartya K. "Poverty: An Ordinal Approach to Measurement." *Econometrica*, March 1976, 44(2), pp. 219-31.
- . "Description as Choice." *Oxford Economic Papers*, November 1980, 32(3), pp. 353-69.
- . "Poor, Relatively Speaking." *Oxford Economic Papers*, July 1983, 35(2), pp. 153-69.
- Zheng, Buhong. "Can a Poverty Index Be Both Relative and Absolute?" *Econometrica*, November 1994, 62(6), pp. 1453-58.

Self-Reliance as a Poverty Criterion: Trends in Earnings-Capacity Poverty, 1975–1992

By ROBERT HAVEMAN AND ANDREW BERSHADKER *

The official U.S. poverty measure identifies those families and individuals with insufficient annual cash income, from either government income support programs or their own efforts, to boost them above a family-size-related, minimum-income threshold, taken to reflect a minimally acceptable level of consumption. Both the conceptual basis and measurement of this indicator have been widely criticized (Patricia Ruggles, 1990; Constance F. Citro and Robert T. Michael, 1995). The annual reported cash income on which the measure relies is but a weak proxy for the family's permanent or potential consumption, in part because it fails to reflect the recipient value of in-kind transfers (e.g., food stamps and Medicaid), the taxes for which the family is liable, the market value of family assets, or the value of voluntary nonwork time. Moreover, the poverty thresholds (the minimum consumption-needs indicators for different family sizes) derive from weak evidence on required consumption levels and equivalence scales. Finally, the official poverty measure rests on the Census Bureau's annual March Current Population Survey, which has been faulted for underreporting income from public transfers, assets, and illegal activities. Hence, the official poverty measure fails to reflect differences among otherwise identical families in *tastes* for leisure (a single-earner two-parent family is more likely to be income poor than an otherwise identical two-earner family); in exogenous *disincentives* to work due to tax-transfer policy differences; and in the generosity of available *public cash benefits*.

Numerous studies have suggested changes in the measure, designed to improve its ability to identify families with insufficient resources to meet consumption needs, including the use of longitudinal data to estimate permanent household income, the substitution of family annual consumption expenditures for cash income, and an expanded definition of net income available for consumption (David Cutler and Lawrence Katz, 1991; Daniel T. Slesnick, 1993; Citro and Michael, 1995; Christopher Jencks and Susan E. Mayer, 1996; David Johnson and Stephanie Shipp, 1998). All of these suggestions accept the underlying social objective on which the official measure rests, namely, that families should be assured some minimal level of living by their own efforts and those of the society.

This philosophical premise, we argue, is less widely accepted now than in the 1960's when the official measure was adopted. Today, concern with "self-reliance" seems paramount in discussions of social and economic policy; to advocates emphasizing this goal, the official measure is but an indicator of failed social policies based on communitarian objectives. Increased public income support, they note, must more than compensate for any associated decrease in individual efforts if official poverty is to be reduced.¹

¹ This new emphasis on self-reliance is consistent with calls for a reduced economic and social-policy role for government. This emphasis is especially obvious in the 1996 welfare-reform legislation, which eliminated the receipt of public transfer benefits by single mothers as an entitlement and imposed firm limits on the period when eligible families could receive support. It is also seen in proposals to privatize Social Security, replace Medicare benefits with medical savings accounts, tighten eligibility criteria for disabled children's receipt of Supplemental Security Income, eliminate public income support for legal immigrants, and replace higher-education grants to students with loans.

* Department of Economics, University of Wisconsin, 1180 Observatory Drive, Madison, WI 53706. Financial support from the Levy Institute of Bard College, the Office of the Assistant Secretary for Planning and Evaluation of the U.S. Department of Health and Human Services, and the Graduate School of the University of Wisconsin-Madison is gratefully acknowledged.

I. Poverty as Inability To Be Self-Reliant

What poverty measure might address the concerns of those who emphasize self-reliance as a social objective? One possibility might be an indicator that reflected the capability of families to meet some minimum level of living by means of their own efforts. Such a poverty measure would reflect the distribution of the capability to be self-reliant in the society. An increase in poverty under this measure would reflect an increase in the proportion of families unable to "get by on their own," and suggest a need to augment the capabilities of these lowest-skilled people. Such a prescription may be more palatable to those concerned with self-reliance than proposals involving the public provision of income support.

One measure of family capability is Gary Becker's (1965) concept of "full income," which includes income realized through market work and the value of leisure time. Adjusting this measure to reflect differences in the size and composition of the consumption unit yields *full income (or potential real consumption) per equivalent consumer unit*. A poverty measure that rests on this concept would indicate a family's ability to support a level of real consumption in excess of needs, that is, to be self-reliant.

Here, we propose such a measure, *net-earnings-capacity (NEC) poverty*. Families that are NEC poor are unable to generate an annual net income stream equal to or greater than their family-specific poverty line, even when the human and physical capital of all adults is fully utilized.

Define gross earnings capacity (GEC) for a family as the level of annual income generated if the head, spouse (if present), and all other prime-aged adults (aged 18–64) worked full-time, full-year (FTFY) at a wage rate reflecting their capabilities and realized the return from their real assets:

$$(1) \quad GEC = EC_H + EC_S + EC_A + \mu$$

where (EC_H) , (EC_S) , and (EC_A) are the earnings capacities of the head, spouse, and other adults in the family, and μ is property income. To take into account exogenous limitations on the full utilization of earnings capacity due to

health problems, disabling conditions, or involuntary unemployment, we adjust the EC values by a factor, Γ , which reflects the time that each individual loses in a year because of these constraints. A second adjustment accounts for expected child-care costs if all adult family members worked FTFY. With these adjustments, net earnings capacity for each family is

$$(2) \quad NEC = (\Gamma_H EC_H + \Gamma_S EC_S + \Gamma_A EC_A + \mu) \\ - (\text{required child care expense}).$$

We obtain EC_i for each working-age adult in the sample from a two-stage procedure. We draw samples of civilian, non-self-employed, nonstudent adults aged 18–64 from each March Current Population Survey (CPS) for 1976–1993, and fit separate models over four race–gender (white–nonwhite; male–female) categories for each year. The first-stage probit equation has FTFY labor-force participation as the dependent variable and is used to compute a selectivity correction term (λ) for each individual. Second-stage log-earnings equations are fit over individuals who have selected into the FTFY work force, with λ added to a vector of individual characteristics. The estimated coefficients from these equations, together with the person's characteristics, yield estimates of EC_i .² These EC_i values are modified to account for the exogenous constraints imposed on individuals by sickness, disability, and the unavailability of jobs. Finally, the modified EC_i estimates [$\Gamma_i EC_i$, in equation (2)] are assembled into family units, to which realized property

² Such a procedure, however, neglects the role of unobserved human-capital and labor-demand characteristics and "luck" in the earnings-determination process and hence leads to an artificially compressed distribution of predicted EC_i for each race–gender group, and for the entire population. To avoid this, we account for unexplained earnings variation within each race–gender group by shocking each estimated EC_i value within a race–gender cell by a random draw from the standard error distribution of the race–gender earnings equations. Further details regarding the sample and estimation are available from the authors upon request.

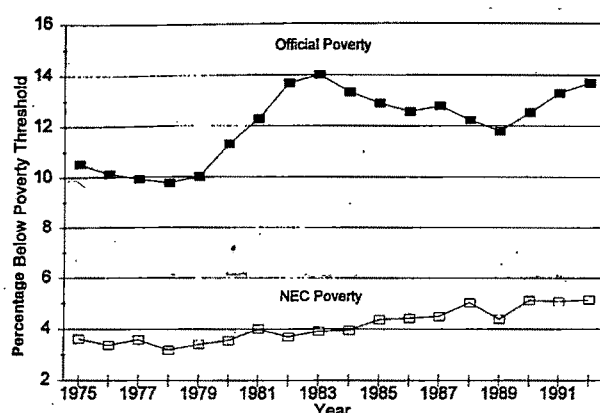


FIGURE 1. NEC VERSUS OFFICIAL POVERTY RATE, 1975–1992, ALL INDIVIDUALS

income (μ) is added, and from which a child-care adjustment is subtracted.³

The resulting NEC value for each family is compared to the official U.S. poverty line for that family. Families for which NEC is less than their poverty line are defined as incapable of being self-sufficient, or as “earnings-capacity poor” (see Haveman and Bershadker [1997] for further details).

II. Trends in U.S. Poverty from 1975 to 1992: Official and NEC Poverty

A. Overall Trends in U.S. Poverty

In Figure 1 and Table 1, we present estimates of both official and NEC poverty rates in the United States from 1975 to 1992. (Detailed numbers underlying these and other estimates are available from the authors upon request.) Over the entire period, there are between 2.5 and 3 times more official poor people than NEC poor. While the absolute gap between the official and NEC poverty rates has widened over time, the percentage change in the NEC poverty index exceeds that of the official measure. Over the period, the official poverty index grew 1.8 percentage points per

³ We assume child care is \$3,000 per child under age 6, and \$1,200 per child aged 6–11, in 1988. We adjust these amounts for inflation in years other than 1988.

TABLE 1—PREVALENCE OF OFFICIAL AND NEC POVERTY

Sample	Poverty (percentage)		
	1975–1977	1990–1992	Trend ^a
A. Official Poverty:			
All individuals	10.19	13.18	1.8
Race of head			
White	6.67	8.27	1.4
Black	27.94	30.86	0.6
Hispanic	21.88	25.98	1.2
Gender of family head			
Male-headed ^b	5.94	7.53	1.6
Female-headed ^c	33.74	29.05	–0.8
Education of head			
Less than high school	20.13	33.55	4.2
High school degree	7.66	13.63	5.2
Some college	5.63	8.30	2.9
College graduate	2.29	2.83	0.8
Family structure			
Married couples ^d	5.45	6.65	1.2
Female headed ^e	43.15	47.23	0.7
White	31.35	34.97	0.9
Black	56.71	58.76	0.2
Hispanic	55.03	57.84	0.2
B. NEC Poverty:			
All individuals	3.52	5.12	3.2
Race of head			
White	1.65	2.48	3.6
Black	13.09	15.10	1.1
Hispanic	9.71	11.97	1.3
Gender of family head			
Male-headed ^b	0.99	1.77	5.7
Female-headed ^c	17.57	14.53	–0.9
Education of head			
Less than high school	7.73	14.40	5.7
High school degree	2.66	5.47	7.3
Some college	1.19	2.72	9.7
College graduate	0.11	0.29	10.0
Family structure			
Married couples ^d	0.94	1.40	4.1
Female headed ^e	26.94	29.34	0.7
White	18.38	21.25	1.0
Black	36.13	36.42	0.2
Hispanic	37.44	38.10	–0.1

^a Trend figures are the coefficients from separate regressions of each indexed poverty series on a time trend. Each indexed series is obtained by setting the poverty percentage for 1975 equal to 100 and projecting the series forward according to the annual percentage changes in the actual series.

^b Single men and men with children, with or without spouse present.

^c Single women and women with children, with or without spouse present.

^d Married couples with or without children.

^e Female heads with children, no spouse.

year.⁴ In contrast, the NEC poverty rate at the end of the period is nearly 150 percent of its initial level, reflecting an index trend of 3.2 percentage points per year. Figure 1 also shows the greater cyclical sensitivity of the official poverty rate compared to the NEC rate; while the official poverty rate reflects actual working time and wage rates, the NEC poverty rate reflects *potential* work time at the implicit predicted wage rate.

The rapid percentage growth in the NEC poverty rate from 1975 to 1992 stands in stark contrast to the findings of Slesnick (1993) and Jencks and Mayer (1996), who report decreases in poverty rates over similar time periods.^{5,6} While their estimates compare family income or consumption *realizations* to a minimum threshold, the NEC indicator compares the *potential* income stream that a family is capable of generating to the threshold. The reductions in the in-

come (consumption)-based poverty rates relative to the increase in NEC poverty suggest that families with the fewest capabilities have increased the utilization of their earnings capacity over this period.

B. Trends in U.S. Poverty among Population Subgroups

Table 1 also summarizes trends in official and NEC poverty for subgroups of the population, identified by the race, gender, and education of the family head, and by family structure. The groups experiencing the most rapid growth in NEC poverty include those in white, male-headed, and married-couple families; such families are not generally considered economically vulnerable. Nevertheless, in spite of large relative increases in NEC poverty for these groups, their NEC poverty rates remained below the overall 5.1-percent national rate by the end of the period.

The most surprising story in Table 1 concerns the groups that have experienced the smallest increases in NEC poverty over the period. Although a high proportion of black or Hispanic families, or those headed by a single female with children, are unable to be self-reliant, these same most-vulnerable groups have experienced decreases (or slow growth) in NEC poverty.

These trends in poverty rates have implications for the composition of the poor population over time. The NEC poverty measure finds that the share of the nation's poor families headed by minorities, single females, and those with low education is larger than that reported by the official measure. At the same time, these vulnerable groups experienced large relative increases in their capacity to earn, and as a result, their share of the NEC poverty population has fallen.

These patterns of change in NEC and official poverty are both revealing and puzzling. Numerous changes in the structure of opportunities, choices in response to these opportunities, and demographic changes have interacted with each other to contribute to these patterns. In Table 2, we list a number of these changes over the 1975–1992 period and summarize their effects on official and NEC poverty rates.

⁴ The statistics used to describe the trends in poverty rates are based on the coefficients from separate regressions of each indexed poverty series on a time trend. Each indexed series is the ratio of each year's poverty rate to the 1975 rate times 100; hence each year's indexed rate is a percentage of the 1975 rate indexed to 100. We interpret the trend coefficient as the average yearly percentage-point change in the indexed series over the 18-year period. While this measure is related to the average annual percentage rate of change of the poverty index, it is not equal to it; both of those measures are sensitive to the absolute value of the initial poverty rate.

⁵ Slesnick (1993) compares consumption expenditures on goods and services (taken to be the indicator of household well-being) to a set of poverty thresholds based on reestimated equivalence scales (constructed using a translog estimation procedure and a large number of household demographic characteristics [see Dale W. Jorgenson and Slesnick, 1987]). He finds that, while both the official poverty measure and his own measure yield a poverty rate estimate of about 12 percent in the early 1970's, the latter poverty rate had fallen to 8.4 percent by 1989, while the official poverty rate had risen to about 14 percent. These poverty estimates and the procedures on which they rest have been critiqued by the U.S. General Accounting Office (1996), among others.

⁶ Jencks and Mayer (1996) calculate a poverty rate for children using an alternative price index (reflecting smaller price-level increases than the official index) and a definition of family income that includes both the income of nonrelatives in the family unit and the value of public in-kind benefits. While the official children's poverty rate rose 5.6 percentage points (from 14 percent to 19.6 percent) from 1969 to 1989, their recalculated poverty rate fell by 1.3 percentage points.

TABLE 2—EFFECTS OF ECONOMIC AND DEMOGRAPHIC CHANGES ON OFFICIAL POVERTY AND NEC POVERTY

Change in economic and demographic patterns	Effect
<i>Official Poverty:</i>	
1) Decrease in real value of public cash transfers	Increase, especially for female-headed families
2) Increase in prevalence of female-headed families	Increase
3) Increase in prevalence of black/Hispanic families	Increase
4) Decrease in prevalence of high-school dropouts	Decrease; perhaps increase in low-education poverty rate
5) Increase in female labor-force participation	Decrease, especially for two-adult families
6) Increase in male joblessness	Increase, especially for two-adult families and males
7) Increase in relative female wage rates	Decrease, especially for female-headed families
8) Absolute decrease in low-skill male wage rates	Increase, especially for two-adult families and males
9) Increase in relative black/Hispanic wage rates	Decrease, especially for racial minorities
10) Increase in within-race-education-gender-group wage inequality	Increase, but muted by low hours worked of low-skilled individuals
<i>NEC Poverty:</i>	
1) Decrease in real value of public cash transfers	No effect
2) Increase in prevalence of female-headed families	Increase
3) Increase in prevalence of black/Hispanic families	Increase
4) Decrease in prevalence of high-school dropouts	Decrease; perhaps increase in low-education poverty rate
5) Increase in female labor-force participation	No effect
6) Increase in male joblessness	No effect
7) Increase in relative female wage rates	Decrease, especially for female-headed families
8) Absolute decrease in low-skill male wage rates	Increase, especially for two-adult families and males
9) Increase in relative black/Hispanic wage rates	Decrease, especially for racial minorities
10) Increase in within-race-education-gender-group wage inequality	Increase

III. Conclusion

Our estimates suggest several conclusions. First, the highest NEC poverty rates are, as expected, found among groups that are generally recognized as the nation's most vulnerable: black, Hispanic, mother-only, and low-education families. The rate of NEC poverty for some of these groups approaches the group-specific official poverty rate, even though the overall NEC poverty rate lies well below the official rate. Second, while both official and NEC poverty rates have increased over the period from 1975 to 1992, growth in the NEC poverty index exceeds that of the official poverty index. Third, in spite of the rapid growth of NEC poverty overall, groups commonly thought of as being the most vulnerable (racial minorities, mother-only households, and those with less education) made progress against NEC poverty over the period. Conversely, groups that are generally viewed as economically secure (whites, two-parent families, and those with relatively high levels of schooling) have experienced *increases* in NEC poverty rates. It is discouraging that even individuals in families with both parents present, and those living in a family headed by someone with some postsecondary schooling, increasingly find themselves unable to escape poverty through their own efforts.

REFERENCES

- Becker, Gary. "A Theory of the Allocation of Time." *Economic Journal*, September 1965, 75(3), pp. 493–517.
- Citro, Constance F. and Michael, Robert T., eds. *Measuring poverty: A new approach*. Washington, DC: National Academy Press, 1995.
- Cutler, David and Katz, Lawrence. "Macroeconomic Performance and the Disadvantaged." *Brookings Papers on Economic Activity*, 1991, (2), pp. 1–61.
- Haveman, Robert and Bershadker, Andrew. "Poverty as 'Inability to Be Self-Reliant': Trends in Earnings Capacity and Official Poverty, 1975 to 1992." Institute for Research on Poverty, University of Wisconsin, 1997.

Jencks, Christopher and Mayer, Susan E. "Do Official Poverty Rates Provide Useful Information About Trends in Children's Economic Welfare?" Mimeo, Northwestern University, 1996.

Johnson, David and Shipp, Stephanie. "Trends in Inequality in the United States Using Consumption Expenditures: The U.S. from 1960-1993." *Review of Income and Wealth*, 1998 (forthcoming).

Jorgenson, Dale W. and Slesnick, Daniel T. "Aggregate Consumer Behavior and Household Equivalence Scales." *Journal of Business*

and *Economic Statistics*, April 1987, 5(2), pp. 219-32.

Ruggles, Patricia. *Drawing the line: Alternative poverty measures and their implications for public policy*. Washington, DC: Urban Institute Press, 1990.

Slesnick, Daniel T. "Gaining Ground: Poverty in the Postwar United States." *Journal of Political Economy*. February 1993, 101(1), pp. 1-38.

U.S. General Accounting Office. *Alternative poverty measures*, GAO/GGD-96-183R. Washington, DC: U.S. General Accounting Office, 1996.

Alternative Historical Trends in Poverty

By DAVID M. BETSON AND JENNIFER L. WARLICK*

The Family Support Act of 1988 called for a scientific review of the official U.S. measure of poverty, reflecting a general dissatisfaction with the current measure, which has not been revised since the mid-1960's. A National Research Council (NRC) Panel of the National Academy of Sciences undertook that review and called for a new approach to poverty measurement.¹ The NRC Panel criticized the current Census Bureau methodology because the current measure of poverty has failed to reflect important economic trends or policies aimed to alleviate the condition it attempts to measure, economic poverty.

This paper presents estimates of how the Panel's recommendations would alter the picture of the number and composition of the poor in the United States over the time period from 1979 to 1994, focusing on trends in poverty among children and the elderly. The current official series shows that the relative gap in poverty rates between children and the elderly has grown since 1979. We present evidence that the poverty gap between children and the elderly is narrowing, not widening, when the Panel's measure of family resources is employed.

I. NRC Panel's Poverty Measure

The concept of economic poverty is simple to state: a family living in poverty lacks goods and services considered essential to human well-being. However, devising an operational measure of poverty that accurately reflects the family's economic well-being in terms of their economic resources has proven difficult.

Families and individuals are officially classified as being in poverty if their annual money

income (including cash earnings, unemployment insurance benefits, cash benefits from other government programs, and other sources of regular nonearned income) before taxes and other deductions falls below official poverty thresholds. The NRC Panel recommended the addition to these resources of several types of family resources that have become major income sources since the 1960's. These include noncash government transfer programs such as Food Stamps, school lunch, and public housing subsidies, and an important cash program administered through the tax code, the Earned Income Tax Credit (EITC).

In the 1960's, the poor were practically exempted from federal income taxation, and very few states taxed low-income families. The only tax they paid on their income was the Social Security Payroll tax at 3 percent of earnings. Today the poor are subject to considerably higher taxes at both the federal level and state level. The NRC Panel recommended that these higher taxes should be taken into account in measuring poverty by subtracting taxes paid from available resources.

Another important trend has been the rapid rise in health-care expenditures directly financed by families (see Gregory Acs and John Sabelhaus, 1995). The NRC Panel recommended that the definition of family resources be altered to reflect the burden imposed by medical expenses by subtracting the amount of medical out-of-pocket spending from the family's available resources.

Finally, any family member who works contributes earnings to the family's resources, but only the earnings net of the cost of employment is available to meet the family's economic needs. The Panel recommended that a limited amount of child-care as well as other work-related expenses should be deducted from the family's available resources.

The NRC Panel also identified three problems with the current specification of the poverty thresholds. First, the current thresholds display an erratic pattern of implicit equiva-

* Department of Economics, University of Notre Dame, Notre Dame, IN 46556.

¹ For a complete description of the Panel's recommendations, see Constance F. Citro and Robert T. Michael (1995). One of us (Betsen) was a member of the NRC Panel on Poverty Measurement and Family Assistance.

lence scales. The Panel recommended that the poverty thresholds be adjusted with an explicit set of scales which would capture the relative needs of families. Second, the current thresholds ignore geographic differences in the cost of living; for example, the cost of housing in New York City is 162-percent higher than in rural Mississippi.

Third, since the poverty measure was first introduced in the United States, there has been no adjustment in the real threshold levels, despite a nearly 30-percent increase in median after-tax incomes of four-person families. The NRC Panel argued for some upward adjustment in the poverty thresholds. More importantly, the Panel argued for a procedure that would update those thresholds over time as spending on necessities such as food, clothing, and housing changes.

II. Alternative Poverty Measures

The NRC Panel's poverty measure would deviate substantially from the current Census definition by reformulating both the definition of the family's available resources and the thresholds against which these resources are to be compared. In this paper, we will focus primarily upon the Panel's recommendations that redefine the family's available resources. Except where noted, we utilize the current set of poverty thresholds.

We compare three alternative definitions of available resources to the current Census definition. The first alternative definition, denoted as the "expanded census" definition, represents the least controversial departure from current practice: adding the value of nonmedical in-kind benefits and the gross amount of EITC received by the family to the current Census money-income definition. The second alternative definition, denoted as "NCR Panel" definition, subtracts the amount of taxes owed, the amount of work-related expenses incurred, and the amount of out-of-pocket spending on health care made by the family, from the expanded Census definition.

While the NRC Panel recommended that the value of public housing provided be included as a resource to the family, it chose not to include the implicit income that families receive from owning their home. Our third alternative

resource definition, "modified NRC Panel," adds the net imputed rent of owner-occupied housing to the family's available resources as defined by the NRC Panel.²

The data we use for our calculations come from the March Supplement of the Current Population Survey for calendar years 1979, 1983, 1989, and 1994. Much of the data needed for the poverty calculations are included on the public-use files provided by the Census Bureau. However, data on child-care and medical out-of-pocket spending were imputed to the data file.³

III. Alternative Poverty Trends

Estimates of the poverty rate for all persons utilizing the four definitions of family resources over the period of 1979–1994 are presented in Table 1. The use of the expanded Census resource definition lowers the overall poverty rate by 19 percent in 1979, 11 percent in 1983, 16 percent in 1989, and 21 percent in 1994. The NRC Panel's resource definition increases the poverty rate of all persons by 22 percent over the official rate in 1979, and by even larger percentage amounts in later years. The addition of net imputed rent from owner-occupied housing lowers the poverty rates compared to the rates derived from the Panel's recommendation but continues to yield poverty rates higher than those produced from the official resource definition. While in 1979 the modified Panel and official Census definitions yield similar poverty rates, there is a 17-percent difference in rates by 1994.

² A previous paper (see Betson, 1995) argued that the value of owner-occupied housing should be limited to the housing portion of the family's needs, which we have assumed to be 30 percent of the family's poverty threshold.

³ Child-care expenses were imputed from regression models estimated from the Survey of Income and Program Participation. Medical out-of-pocket spending was based upon regression analysis of the 1987 National Medical Expenditure Survey. See Patricia Doyle (1997) for an evaluation of the imputation procedure for medical out-of-pocket spending that is used in this paper. Annual work-related expenses were assumed to be \$750 for a full-year worker in 1992 and were prorated by the number of weeks reported worked by each individual in the family.

TABLE 1—IMPACT OF RESOURCE DEFINITION ON THE POVERTY RATES OF ALL PERSONS (PERCENTAGES)

Definition	1979	1983	1989	1994
Census	11.7	15.2	12.8	14.6
Expanded Census	9.5	13.6	10.7	11.6
NRC Panel	14.3	19.6	17.4	18.8
Modified Panel	11.9	16.5	15.2	17.0

The alternative resource definitions yield different pictures of the sensitivity of poverty to recession and recovery. From 1979 to 1983, the official poverty rate increased by 30 percent. Over the same time period, the three alternative resource definitions indicated significantly larger percentage increases in poverty rates. As the economy recovered between 1983 and 1989, the official poverty rate fell by 16 percent. The poverty rate under the expanded Census resource definition fell by 21 percent, but the recovery was reflected less by both the NRC Panel and modified Panel definitions which fell only 11 percent and 8 percent, respectively. Between 1989 and 1994, the official poverty rate rose by 14 percent, while all the other measures showed slightly less sensitivity to the declining economy.

Although the level of poverty in the total population is an important social indicator, equally important to our thinking about poverty are the poverty rates of various subgroups in the population. Table 2 presents the poverty rates of children and the elderly using the four resource definitions.

Between 1979 and 1994, the official child poverty rate rose from 16.4 percent to 21.8 percent. In 1979, the poverty rate for the elderly was 15.2 percent; by 1994, it had fallen to 11.7 percent. These trends are consistent with an ever-widening gap in the incidence of poverty among children and the elderly. In 1979, the child poverty rate was 8-percent higher than the poverty rate of the elderly; but by 1994, poverty among children was 86-percent higher.

Using the expanded Census resource definition, the poverty rates of both children and the elderly are lower than the currently published poverty rates. However, the historical trend in child poverty rates relative to the el-

TABLE 2—THE IMPACT OF ALTERNATIVE RESOURCE DEFINITIONS ON CHILD AND ELDERLY POVERTY RATES (PERCENTAGES)

Definition	1979	1983	1989	1994
Child Poverty Rates:				
Census	16.4	22.3	19.6	21.8
Expanded Census	12.6	19.7	16.0	16.6
NRC Panel	18.2	27.4	23.7	25.3
Modified Panel	15.7	23.9	22.0	23.6
Elderly Poverty Rates:				
Census	15.2	13.8	11.4	11.7
Expanded Census	13.5	12.0	9.6	9.7
NRC Panel	22.4	20.4	19.4	20.9
Modified Panel	16.2	14.6	15.0	16.3

derly is consistent with the widening poverty gap between children and the elderly found in the official series. In 1979, the child poverty rate was 7-percent lower than the poverty rate of the elderly; but by 1994, poverty among children was 71-percent higher.

Between 1979 and 1983, the NRC Panel, modified Panel, and Census resource definitions produce similar trends in poverty rates: increasing poverty among children with declining poverty in the elderly population. The relative poverty gap widens over this four-year period. After 1983, a different trend emerges. In 1983, the child poverty rate using the Panel definition was 34-percent higher than the poverty rate of the elderly, while the modified Panel definition yielded a 64-percent differential in poverty rates. By 1994, the differential in poverty rates fell to 21 percent under the Panel's definition and to 45 percent with the modified Panel resource definition. Between 1983 and 1994, the poverty gap narrows. We find that the diminishing gap in the incidence of poverty between children and the elderly is attributable to the differences in medical out-of-pocket spending between families with children and the elderly. The poverty rates of both groups increase when these expenses are subtracted from income; however, the effect is much greater for the elderly.

IV. Setting the Poverty Thresholds

Previous research has shown that the Panel's recommendations for determining the poverty thresholds can have a differential impact on the poverty rates of various subgroups in the population. For example, a previous paper found that the poverty rate of single-elderly individuals would be greatly reduced by the use of the Panel's equivalence scales, while poverty among children would be unaffected (see Betson, 1996).

To quantify the sensitivity of our findings of a diminishing poverty gap between children and the elderly to choice of poverty thresholds, we identified new poverty populations by uniformly increasing and decreasing the current poverty thresholds by 20 percent. We then computed the ratio of the percentage change in the poverty rate of children to the percentage change in the poverty rate of the elderly. If this ratio is greater (less) than 1 when we lowered poverty thresholds, the poverty gap would further narrow (widen) with a choice of a lower threshold. Conversely, if the ratio is less (greater) than 1 when we raised the poverty thresholds, the poverty gap would narrow (widen). The average ratio over the four years of data when using the Panel's resource definition was 1.09 when the thresholds were lowered and 0.92 when the thresholds were raised. When the modified Panel definition was employed, the average ratio was 1.09 when the thresholds were lowered and 0.82 when the thresholds were raised. These results suggest that our findings of a diminishing poverty gap between children and the elderly will not be altered by alternative definitions of poverty thresholds that maintain the current equivalence scales.

V. Conclusion

In this paper, we have provided evidence that how we measure a family's resources is an important factor in determining the number

and the composition of America's poor. The current resource definition employed by the Census Bureau is commonly acknowledged to be inadequate for measuring poverty. We have shown that incremental corrections that account for the value of in-kind benefits and the EITC do not fully reflect the resources that a family has to meet its nonmedical necessities. Only if the measurement of the family's resources also takes into account the taxes paid by the family, the cost of earning one's income, the value of home-ownership, and the medical expenses incurred by the family, will an accurate ordering of the nation's families be achieved. Moreover, less-comprehensive measures distort the trends in the rates of poverty among children relative to those among the elderly. The more-comprehensive resource measure envisioned by the NRC Panel and modified here shows that, contrary to the current poverty series, the relative poverty gap between children and the elderly is narrowing, not widening.

REFERENCES

- Acs, Gregory and Sabelhaus, John. "Trends in Out-of-Pocket Spending on Health Care, 1980-92." *Monthly Labor Review*, December 1995, 118(12), pp. 35-45.
- Betson, David M. "The Effect of Home Ownership on Poverty Measurement." Mimeo, University of Notre Dame, December, 1995.
- _____. "Is Everything Relative? The Role of Equivalence Scales in Poverty Measurement." Mimeo, University of Notre Dame, 1996.
- Citro, Constance F. and Michael, Robert T. *Measuring poverty: A new approach*. Washington, DC: National Academy Press, 1995.
- Doyle, Patricia. "How Do We Deduct Something We Do Not Collect? The Case of Out-Of-Pocket Medical Expenditures." Mimeo, U.S. Bureau of the Census, Washington, DC, 1997.

Poverty-Measurement Research Using the Consumer Expenditure Survey and the Survey of Income and Program Participation

By KATHLEEN SHORT, MARTINA SHEA, DAVID JOHNSON, AND THESIA I. GARNER*

The most recent comprehensive examination of poverty measurement in the United States was conducted by the National Academy of Sciences (NAS) Panel on Poverty and Family Assistance (Constance F. Citro and Robert T. Michael, 1995). In their report, the Panel recommended changing the definition of both the poverty thresholds and the resources that are used to measure poverty. In this paper we implement a number of the Panel's basic procedures, with slight modifications, to obtain experimental poverty rates for 1991 to 1996.¹

This paper presents poverty estimates using thresholds derived from the 1989–1991 Consumer Expenditure Survey (CEX), and using family resources based on the 1991 panel of the Survey of Income and Program Participation (SIPP) and the March 1992 Current Population Survey (CPS). The resulting experimental poverty rates are compared to those based on the official measure. While most previous work has examined the new poverty measure exclusively using the CPS, this paper presents, for the first time, estimates from the SIPP, the survey that the Panel recommended should become the official source of poverty resource measurement. Additional estimates from the CPS from 1992–1996 are presented in order to examine the behavior of these experimental poverty rates over time.

Our findings reveal that changes in the poverty rates based on the official and the experimental measures are similar over time. We show that poverty rates using SIPP data are below those using the CPS. We also show that using the experimental poverty measure yields a poverty population that looks more like the total population in terms of various demographic and socioeconomic characteristics than does the poverty population based on the current official measure.

I. Defining the Thresholds

The Panel recommended that the poverty thresholds should be based on a percentage of the median expenditures for a basic bundle which includes food, clothing, shelter, and utilities. A small multiplier is applied to this value to account for other needs (e.g., household supplies, personal care). The actual expenditures of a consumer unit, comprising two adults and two children, from the CEX data are used. Following Garner et al. (1998) we use the average of upper and lower values for the percentages and multipliers to obtain a poverty threshold for the reference unit (shown in Table 1). The resulting threshold is very close to median expenditures on the basic bundle. We update the threshold from 1991 using the CPI-U. While the Panel recommended updating by the change in median expenditures each year, Johnson et al. (1997) showed that the change in median expenditures were similar to the inflation rate over this entire period, but the annual changes were more volatile than the inflation rate.

The reference threshold is adjusted to reflect geographic differences in costs, using inter-area housing price indexes based on data from the 1990 Census on gross rent for apartments (as did the Panel). We use a two-parameter equivalence scale that accounts for the differ-

* Short and Shea: Housing and Household Economics Statistics Division, Census Bureau, Washington, DC 20233; Johnson and Garner: Division of Price and Index Number Research, Bureau of Labor Statistics, Washington, DC 20212. All views expressed in this paper are those of the authors and do not necessarily reflect the policies of the Bureau of the Census or the Bureau of Labor Statistics.

¹ For a more comprehensive version see:

<http://www.census.gov/www/hhes/povmeas.htm>

TABLE 1—POVERTY THRESHOLDS FOR UNITS COMPRISING TWO ADULTS AND TWO CHILDREN

Year	Poverty threshold (\$)	
	Official	Experimental
1991	13,812	13,891
1992	14,228	14,309
1993	14,654	14,738
1994	15,029	15,115
1995	15,455	15,543
1996	15,911	16,002

ing needs of adults and children and the economies of scale of living in larger families. This scale is $(A + PC)^F$, where A and C represent the number of adults and children, P represents the adult-equivalent of one child, and F represents the economies of scale. We use $P = 0.7$ and a scale economy factor $F = 0.65$, since these scales minimize the effect on overall poverty and are most similar to the current scales; however, different equivalence scales can change the composition of poverty (see Citro and Michael, 1995; Johnson et al., 1997).

II. Defining Resources

Following the Panel's recommendation, we use an experimental resource measure that is based on money income plus the value of in-kind transfers but which excludes taxes, child-support paid, and work-related expenditures. We use the following in-kind valuation methods:

- (i) Food stamps, reported face value;
- (ii) School lunch, reported receipt, imputed value;
- (iii) School breakfast, reported receipt, imputed value in SIPP;
- (iv) Housing subsidies, Department of Housing and Urban Development's fair-market rents in SIPP (see Martina Shea et al., 1997), modeled in the CPS (see Census Bureau, 1997);
- (v) Energy assistance, reported in SIPP;
- (vi) Women, Infants, and Children Program (WIC), reported in SIPP.

Benefits from WIC, school-breakfast, and energy-assistance programs are added to the SIPP resource measure but not the CPS measure. Not including these three benefits increases the standardized experimental poverty rate by 0.2 percentage points in the SIPP measure.

From the cash and in-kind transfer total we subtract the following expenses:

- (i) Work-related transportation and miscellaneous expenses, a fixed amount per week per working adult, not to exceed earnings;²
- (ii) Child-care expenses, reported in SIPP and imputed to the CPS (see Short et al., 1996);
- (iii) Medical out-of-pocket expenditures, imputed in both surveys (see David Betson, 1997a, b; Patricia Doyle, 1997);
- (iv) Taxes, imputed in the CPS;
- (v) Child-support paid, reported in SIPP.

Our treatment of the last two elements differs between the two surveys. In the CPS, taxes paid are modeled in every year, including the value of the Earned Income Credit (EIC) received. The SIPP collects information on taxes paid in an annual tax module; we are currently evaluating these data to develop a tax estimation procedure for the SIPP. For the purpose of this paper, we do not subtract taxes from income for the SIPP analysis. Our calculations show that accounting for taxes in our standardized experimental CPS measure increases the poverty rate by about 1.0 percentage point. Further, information on child-support payments are not available in the CPS and, therefore, are not included in the CPS estimates reported here. Calculations show that accounting for child-support paid in the SIPP experimental measure increases the poverty rate by less than 0.1 percentage point.

III. Results

At this stage of analyzing the Panel's recommendations, poverty rates are important

² The Panel estimate of \$14.42 for 1992 was price-adjusted for other years.

TABLE 2—POVERTY RATES, 1991 (PERCENTAGE POOR)

Sample	NAS experimental measure					
	Official definition		Standardized		Nonstandardized	
	CPS	SIPP	CPS	SIPP	CPS	SIPP
All persons	14.2	12.1	14.2	14.2	18.9	13.6
Children	21.8	19.6	19.9	20.0	26.4	18.9
Elderly	12.4	9.0	14.9	15.3	20.3	14.5
White	11.3	9.3	12.1	12.0	16.1	11.5
Black	32.7	29.0	27.4	28.4	36.7	26.8
Hispanic	28.7	27.6	30.6	30.8	40.0	29.5
One or more workers	9.3	6.6	10.4	9.6	14.3	9.0
Married-couple families	7.2	6.3	8.3	9.3	11.9	8.8
Female-householder families	39.7	35.5	35.7	35.2	45.0	33.6

as a starting point from which to examine trends and the composition of the poverty population. In Table 2, poverty rates using the official thresholds and resource measure for different demographic groups are compared to the poverty rates based on our implementation of the Panel's proposed method using the SIPP and CPS for 1991. As shown, poverty rates using the official definition with SIPP data are smaller than official CPS-based poverty rates.³ In order to examine the effects on the composition of the poverty population, we adjust the experimental thresholds by a percentage of the threshold to obtain an overall poverty rate equal to the official rate. The standardized rates in Table 2 show that children, blacks, and people in female-householder families are less likely to be classified as poor under the new measure, while all other groups shown are more likely to be classified as poor.

Since the experimental standardized poverty rate is lower than the official rate for children, blacks, and persons in female-householder families, we would expect that their representation in the poverty population would be lower, and vice versa for those with higher

³ SIPP was designed to do a more complete job of collecting income data than the CPS.

TABLE 3—PROPORTION OF THE POPULATION BY CHARACTERISTIC, 1991 (PERCENTAGES)

Sample	Poverty population					
	Total population		Official definition		NAS experimental measure, standardized	
	CPS	SIPP	CPS	SIPP	CPS	SIPP
Children	26.2	26.9	40.2	43.9	36.8	37.7
Elderly	12.2	11.5	10.6	8.6	12.8	12.4
White	83.7	83.3	66.5	64.2	71.2	70.3
Black	12.5	12.5	28.7	30.1	24.1	24.9
Hispanic	8.8	9.3	17.8	21.2	18.9	20.0
One or more workers	84.5	82.0	54.9	45.1	61.7	55.2
Married-couple families	79.7	80.3	44.8	44.7	50.6	54.2
Female-householder families	16.4	16.6	50.9	52.2	44.4	42.4

rates. As seen in previous research, using the new measure results in a poverty population that more closely resembles the total population. This is illustrated in Table 3, which shows the composition of the total population versus that of the poverty population under the different measures.

Finally, Table 4 shows that over the 1991–1996 period, rates under the official and experimental methodologies behave similarly, increasing over the 1991–1993 period and decreasing over the 1993–1996 period. The table shows standardized experimental poverty rates calibrated to the 1991 official rate. The official rate rises from 14.2 percent to 15.1 percent from 1991 to 1993 and falls to 13.7 percent by 1996.⁴ The standardized experimental rate rises from 14.2 percent to 15.4 percent from 1991 to 1993 and falls to 13.4 percent by 1996. However, over the 1993–1996 period, poverty rates drop more under the experimental measure for some

⁴ Controlling the experimental measure to match the official poverty rate in 1996 instead of 1991 shows the same pattern: an increase from 14.5 percent in 1991 to 15.7 percent in 1993, followed by a reduction to 13.7 percent in 1996.

TABLE 4—POVERTY RATES, 1991–1996, CPS
(PERCENTAGE POOR)

Sample	1991	1992	1993	1994	1995	1996
<i>Official Measure:</i>						
All persons	14.2	14.8	15.1	14.6	13.8	13.7
Children	21.8	22.4	22.7	21.8	20.8	20.5
Elderly	12.4	12.9	12.2	11.7	10.5	10.8
White	11.3	11.9	12.2	11.7	11.2	11.2
Black	32.7	33.4	33.1	30.6	29.3	28.4
Hispanic	28.7	29.6	30.6	30.7	30.3	29.4
One or more workers	9.3	9.7	9.9	9.6	9.5	9.5
Married-couple families	7.2	7.7	8.0	7.4	6.8	6.9
Female-householder families	39.7	39.0	38.7	38.6	36.5	35.8
<i>NAS Experimental Measure (Standardized):</i>						
All persons	14.2	15.0	15.4	14.3	13.4	13.4
Children	19.9	20.7	21.0	19.3	17.8	17.7
Elderly	14.9	16.1	16.2	15.7	14.6	15.2
White	12.1	12.7	12.9	12.3	11.5	11.5
Black	27.4	29.2	30.1	25.2	24.6	24.5
Hispanic	30.6	31.1	31.0	30.1	28.0	27.8
One or more workers	10.4	10.7	11.1	10.2	9.7	9.7
Married-couple families	8.3	9.1	9.2	8.3	7.6	7.6
Female-householder families	35.7	35.3	35.2	34.1	31.3	31.6

groups, such as children and blacks. This drop appears to be due to the addition of the Earned Income Credit in the resource measure. This result highlights the ability of the new measure to capture the effects of many tax and transfer policies.

IV. Next Steps

Future poverty-measurement research will address refinements in the thresholds and the way in which resources are defined. Further work on the threshold side includes examining other geographic adjustments and equivalence scales. While the procedure used here to adjust for geographic differences in housing prices is understandable and operationally feasible, it does not account for housing cost differences within areas or for quality differences. Additionally, since the choice of an equivalence scale can have large effects on the composition

of the poverty population, the selection of appropriate equivalence scales must be further examined.

On defining resources or "income," the largest remaining challenge involves calculating taxes to arrive at an after-tax income measure in the SIPP. Other work involves further examination of the imputation procedures used to produce the medical out-of-pocket and housing-subsidy values. Because aggregate imputed values are calibrated to benchmark totals, outcomes are quite sensitive to changes in these totals. Finally, new data collected in the SIPP this year, which allows a statistical match to the Medical Expenditure Panel Survey, may result in improved methods of valuing this element of the measure of resources.

REFERENCES

- Betson, David. "Imputing Medical Out-of-Pocket Expenditures from NMES Data." Unpublished manuscript, University of Notre Dame, December 1997a.
- . "In Search of an Elusive Truth, 'How Much Do Americans Spend on Their Health Care?'" Unpublished manuscript, University of Notre Dame, 7 April 1997b.
- Census Bureau. *Poverty in the United States: 1996*. Washington, DC: U.S. Government Printing Office, September 1997.
- Citro, Constance F. and Michael, Robert T. *Measuring poverty: A new approach*. Washington, DC: National Academy Press, 1995.
- Doyle, Patricia. "How Do We Deduct What We Do Not Collect?" in *Proceedings of the Government and Social Statistics Section*. Alexandria, VA: American Statistical Association, August 1997, pp. 38–47.
- Garner, Thesia I.; Shipp, Stephanie; Paulin, Geoffrey; Short, Kathleen and Nelson, Charles. "Poverty Measurement in the 1990s." *Monthly Labor Review*, 1998 (forthcoming).
- Johnson, David; Shipp, Stephanie and Garner, Thesia I. "Developing Poverty Thresholds Using Expenditure Data," in *Proceedings of the Government and Social Statistics Section*. Alexandria, VA: American Statistical Association, August 1997, pp. 28–37.

Shea, Martina; Naifeh, Mary and Short, Kathleen.

"Valuing Housing Subsidies in a New Measure of Poverty: SIPP," in *Proceedings of the Government and Social Statistics Section*, Alexandria, VA: American Statistical Association, August 1997, pp. 376-81.

Short, Kathleen; Shea, Martina and Eller, T. J.

"Work-Related Expenditures in a New Measure of Poverty." Unpublished manuscript presented at the annual meeting of the American Statistical Association, Chicago, IL, August 1996.

AFRICAN-AMERICAN ECONOMIC GAINS: A LONG-TERM ASSESSMENT[†]

Race and Class in Postindustrial Employment

By GERALD D. JAYNES*

The decline in black men's employment since 1940 has posed a significant puzzle. Black men's 1940 employment rate (76 percent) exceeded white men's by 2 points, but by 1960 their 69-percent rate was 7 points below white men's. By 1985, the black rate of 56 percent was 14 points below whites'. The racial employment gap exists for all significant subgroups but is largest at low skill/experience levels. Explanations abound, but most commentators endorse some combination of discrimination, declining demand for less-skilled inexperienced workers, changing values toward work among blacks, government welfare, and suburbanization of employment. In this maze even economists have admitted that the problem may transcend ordinary methods of analysis. Albert Rees (1986 p. 623) concluded that racial employment differences may stem from aspects of the "contemporary social structure or culture of the black ghetto" making it "hard to disentangle causes from effects" so that "the techniques of economics may be less useful than those of social anthropology."

One need not abandon economics, but as Rees suggests, techniques of social anthropology can clarify long-term changes in black employment. The parsimony of the choice-theoretic framework of modern economics produces what the anthropologist Clifford Geertz (1973 p. 27) has termed "thin description"; economic models pro-

vide a schematization of theoretical possibilities that generally cannot be finely distinguished by the statistical data available. "Thick description," by centering on the subjects' personal views of the motivations, attitudes, and beliefs undergirding their observed behavior, adds flesh to these schematizations. It constructs interpretations derived from the meaning social actions have for the actors performing the acts. Such interpretations contextualize the actors' conceptual meanings within the social structure in which their actions occur, and this can help the analyst disentangle the cultural factors from structural determinants of behavior.

Thick description interprets the conceptual meaning of concepts such as the reservation wage and opportunity cost to determine why and how decisions are made in terms respectful to the decision-maker's perspective. In effect it asks: what are the arguments of the agent's preference field? To do this subject-centered analysis requires discursive data from its subjects. Recent studies of price and quantity decision-making based on surveys of actual decision-makers (Truman Bewley, 1997; Alan S. Blinder et al., 1998) and the NBER employment survey of inner-city black males (Richard B. Freeman and Harry J. Holzer, 1986) are examples of such discursive data. Yet, as valuable a resource as the NBER sample has proved to be, it has limitations. Like all survey instruments, its design was constructed by researchers who solicited data from respondents. The researcher and subject interact to produce the conceptual context surrounding the data. Since a good design will be driven by theory, questions crucial to eliciting the respondents' own understanding of their behavior may be omitted. Survey instruments must be supplemented with unsolicited data:

[†] *Discussants:* Bernard E. Anderson, U.S. Department of Labor; Kaye G. Husbands, Williams College; Donald R. Williams, Kent State University.

* Department of Economics, Yale University, Box 1972, Yale Station, New Haven, CT 06520. All numbers are calculated from unpublished Census data, available from the author upon request.

discursive data that is independent of the researcher.

I am currently combining choice-theoretic theory and solicited data (e.g., Census data and ethnographic research) with unsolicited data (autobiography, letters, primary historical documents, and a subgenre of rap music I label "testimonial rap") to construct subject-centered analyses of the behavior of the urban poor. This paper gives an example of this methodology by applying it to the question of black employment.

I. Some Facts

Table 1 shows that in 1940, compared to other regions, the high black employment rate in the South was an anomaly; black employment in the South converged to the other regions by 1960, and therefore, the decline in black employment during the two decades was a Southern adjustment. The decline in the South accompanied black urbanization, a phenomenon attributed to the push factor of agricultural mechanization (John Cogan, 1982). In a purely descriptive sense, the decline in black employment was attributable to the decline in the demand for agricultural labor. However, careful consideration reveals that this description does not reveal much. The decline in white male employment was also a Southern phenomenon driven by urbanization. Black men's adverse relative employment in 1960 was due to preexisting employment problems in cities. The crucial problem is not explaining why black employment fell but explaining why urban black employment was and remains so low.

Several observations merit emphasis. First, the data (not shown) demonstrate that employment and black-white differences among fixed education-age groups changed little between 1940 and 1970 (teenagers and those at or near retirement ages were exceptions). Thus changes in social-welfare programs of the 1960's cannot be important primary causes. Also, since with urbanization Southern black employment converges to the non-South, North-South differences in values were not primary causes. The differences must be rural-urban. Finally, convergence occurs so quickly that the causes must relate to

TABLE 1—BLACK MALE EMPLOYMENT PERCENTAGES BY REGION

Region	1940	1960	1985
Northeast	64	72	50
North Central	67	68	47
West	65	74	59
South	79	69	58
United States	76	69	56
Black/white	1.02	0.90	0.79

Source: U.S. Bureau of the Census.

changes in social structure, not cultural values. A conclusion consistent with the observation that the most significant black-white differences are post-1970, a period when real wages of less-skilled men fell drastically, is that all less-skilled men reduced their labor supply. Below I focus on why black employment levels fell more.

II. Identity And Labor Supply

Crucial to my interpretation of the determinants of labor supply is the interaction of culture and social structure in the formation of an individual's social identity and worldview. Understanding a worldview and its relationship to social identity requires subject-centered analysis. Space limits the discussion to one example, but it is informative. Lower-working-class blacks' use of the metaphor "slave" to describe certain jobs spans the 20th century manifesting considerable continuity in worldview. To see the relation between worldview, identity, and labor supply requires explication of the social meaning of "slave." As metaphor "slave" does not refer to all employment. The characteristics of a "slave" are demeaning and close supervision, low pay, and little opportunity for career advancement. Use of the term is not restricted to hustlers or to men. New York's black domestics referred to the corner of 174th Street and Sedgwick Avenue, where whites would drive by and hire workers, as the "Bronx slave market."

To hold a slave job is likely to damage one's self-esteem and identity. Thus, a slave is something to be avoided. As blacks urbanized, the interaction of culture and changing social structure significantly altered three determi-

nants of labor supply: the proportion of jobs viewed as slave jobs increased, as did the net benefits of feigning deference to whites; real and perceived blocks to opportunity decreased the means of avoiding slave jobs; and identity-formation among the young crystallized around new role models for maintaining self-esteem. These changes were interactive.

III. Avoidance and Self-Respect: Equilibria in Racialized Labor Markets

Prior to the 1970's, mores regulating interracial contacts, although less blatant and rigid in the North, included an expectation that blacks would assume deferential roles. Blacks interacting with white employees and employers were forced to submit to demeaning conditions of work that transcended the usual kinds of differential treatment in hiring, job assignments, and promotions currently understood to be employment discrimination. Blacks' major defense against such relations was avoidance of whites. Given alternatives, they avoided working with or for whites unless they were compensated for the psychological costs incurred.

The complex issues of race and class underlying treatment of blacks are considerably more complicated today. Many blacks have achieved levels of success at virtually all levels of the occupational structure. Yet, blacks say that, on the job and elsewhere, they are often subjected to rude treatment, excluded from social networks, or merely tolerated by belligerent or condescending whites assuming superior intellectual capabilities. The costs of interracial contact continue to induce blacks of all class backgrounds to adopt strategies designed to avoid interracial contacts with high probability of generating a loss of self-esteem. Several studies document black professionals in predominately white organizations as being discontented, under great stress, and consequently fleeing their fields in considerable numbers to find environments more insulated from demeaning interracial contacts. A common complaint is that, to smooth relations with white coworkers, and therefore have a chance at a successful career, blacks must mask their views on race and perfect a persona that downplays black identity (Edward Irons and Gilbert

Moore, 1985; Ellis Cose, 1993; Lawrence D. Bobo, 1997).

In such labor markets, blacks' reservation wages will include the expected costs of interacting with whites. Insufficient wages will not be accepted, or if the psychic costs prove higher than expected, blacks will quit or cause enough "trouble" to be fired. Black workers should be expected to undergo significant job search and have higher turnover rates than whites. Employers who do not discriminate in the commonly accepted meaning of the term may be no better at providing blacks a "comfortable" work environment because a primary impediment is lack of knowledge. Moreover, providing such an environment will raise costs (especially if a trade-off between black and white comfort exists). Equilibrium may entail an extreme form of the segregation predicted by the Becker-Arrow model of discrimination (Gary S. Becker, 1957; Kenneth J. Arrow, 1972, 1973): in the absence of a cadre of black entrepreneurs, blacks may exit the labor market for earning activities that minimize their relations with whites. Because there is no a priori reason to think that the psychic costs of black-white relations differ by class, and because there is reason to believe that racial climates on less-skilled jobs may be less gentile, lower-class blacks who will have lower wage offers will be most likely to leave the market. In equilibrium, some blacks will be employed in the interracial labor market, and some will be employed or self-employed in a black market.

Because black-white residential segregation had very different employment effects in rural and urban settings, urbanization increased on-the-job interracial contacts, thereby multiplying the proportion of jobs classified as slave jobs. In rural economies, places of work and residence coincide; residential segregation insulated blacks from both competition and employment contact with whites. In urban economies, blacks faced direct employment competition and contact with whites because high population densities in compact land areas accommodated residential segregation but were incongruent with absolute employment segregation.

In agriculture, blacks' demand for self-employment to avoid contact with whites

led to the sharecropping tenancy system, which permitted blacks a large degree of independence from close supervision and contact with whites. At mid-20th century, blacks continued to view farming as a means of avoiding working under the close supervision of whites who talked to them with extreme disrespect (Charles S. Johnson, 1943; R  ger L. Ransom and Richard Sutch, 1977; Jaynes, 1986). In urban economies economic segregation was occupational. Blacks were segregated into the most menial jobs where pay was meager and employment often very unstable. Yet discrimination and occupational segregation did not buffer blacks from close work contact with whites who expected blacks to assume subordinate and demeaning roles. Racist attitudes among whites restricted the occupational choices of black women to stable but extremely low-paying, physically demanding and psychologically demeaning domestic service. Blacks avoiding white contact worked in the black segregated economy or became self-employed.

IV. Class Structure and Urban Opportunity

Black class structure circa 1940 illustrates the incentives for youths to enter extra-legal activities. Each class contained members occupied in gray- or black-market activities, but their proportion declined as class status rose. Even upper-class "respectables" accorded status to wealthy entrepreneurs like Casper Holstein, a West Indian immigrant porter who rose to preside over Harlem's busiest and most prosperous business, the illegal numbers or policy game. Prestigious role models like Holstein underscore Gunnar Myrdal's observation that many lower-class urban youths might aspire to a career in vice and crime (St. Francis Drake and Horace Cayton, 1945; Myrdal, 1962 p. 705; David Levering Lewis, 1981). In contrast to legal work where blacks were largely proscribed to menial jobs, a sharp, hardworking youth could aspire to a criminal career.

The numbers game was a major source of gambling, and a thriving business and source of employment for thousands of blacks. Many of the black men imputed zero earnings by the Census (22 percent in 1959) could have

been found in this or other illegal industries. In Harlem during the mid-1950's numbers runners (entry-level employees) earned approximately the annual earnings of the median full-time black male in New York City. Given the risks involved (see Alex Haley, 1965) and the comparability of entry-level earnings to earnings of black men generally, the young did not enter with the prospects of immediate wealth. Many discounted or ignored the risks and entered criminal careers hoping to attain wealth and ghetto fame. Equally important, they worked outside the formal economy, avoiding the slave jobs and their demeaning race relations.

Those involved in the extralegal lifestyle carved out a social existence structurally similar to sharecropping. The possibility of attaining success, landownership in the rural south or wealth and prestige as a ghetto hustler, and the open access to the "chase" provided the means by which young people could maintain an identity and self-esteem by pursuing a life plan that they could value. The institutions were similar in another respect: while eager young entrants to the "chase" were confident they would defy the odds against success, most hustlers and criminals, similar to most tenant farmers who found themselves landless and poor at early middle-age, would be disappointed, learning that crime usually leads to lengthy periods of incarceration or to an abrupt and violent end to a short life.

The social structures of inner cities in the 1940's and 1990's are similar with two major exceptions: the numbers economy has been replaced by the drug economy, and middle-class blacks are a much lower presence. Continuities in worldview suggested by the persistent social meaning of "slave" support other methods of thick description in indicating that, for the less skilled, labor-supply decisions remain similar. Ethnographical and autobiographical accounts of life in cities confirm this motivation for high labor turnover rates and extralegal hustling. Employee-employer relations across the color line remain a major impediment to labor-force participation among less-skilled blacks (Elijah Anderson, 1980; Freeman and Holzer, 1986 pp. 16-18; Joleen Kirschenman and Kathryn Neckerman, 1991).

A New Yorker of the 1950's explains why he quit his "slave" job: "My brothers, they can't stand to be around gray people [whites]. That's why they all stand around 143rd Street and take numbers. I guess we couldn't make it outside of some Harlem somewhere. We weren't cut out to play that boy role." He continued: "A cat got to take a whole lotta s--t for fifty dollars a week." "I don't think the stuff that a man has to take down here is worth fifty dollars a week; it's worth a lot more, at least ten times more" (Claude Brown, 1965 p. 296). Nathan McCall, who sold drugs in Portsmouth, Virginia, during the 1970's, described similar motivations: "All that mattered to me was that he chose not to earn his living slaving at the shipyard or bowing to white folks on some other gig. Turkey had an older brother who was definitely a hustler. He and his friends gave me my first inside look at those blacks who live and operate almost completely outside the white man's system. They make money doing business with other blacks" (McCall, 1994 pp. 80-81).

Access to means of avoidance and its costs are highly differentiated by class. Middle-class African-Americans working in majority-white organizations are compensated generously for any stress and loss of self-esteem they might suffer. Moreover, they possess superior self-employment options. Less-skilled blacks frequently face more blatant affronts to their self-worth for remuneration that leaves them below or just above poverty. For many, the declining-wage, post-1970 labor market has not met their price for adjustment and conformity. Given the context of contemporary race relations and post-1970 wage deterioration (real mean weekly wages of black dropouts declined 32 percent during 1969-1984), a gap exists between blacks' offered and reservation wages. Given their limited skills, avoidance of a slave job leads to rebellion and crime or, for young women, public assistance.

Many less-skilled younger men find their identity and self-esteem threatened in the labor market. They enter the underground economy where, in the short term, they suffer less embarrassment and psychological fear.

REFERENCES

- Anderson, Elijah. "Some Observations of Black Youth Employment," in Bernard E. Anderson and Isabel V. Sawhill, eds., *Youth employment and public policy*. Englewood Cliffs, NJ: Prentice Hall, 1980, pp. 64-87.
- Arrow, Kenneth J. "Some Mathematical Theories of Race in the Labor Market," in Anthony Pascal, ed., *Racial discrimination in economic life*. Lexington, MA: Heath, 1972.
- . "The Theory of Discrimination," in Orley Ashenfelter and Albert Rees, eds., *Discrimination in labor markets*. Princeton, NJ: Princeton University Press, 1973.
- Becker, Gary S. *The economics of discrimination*. Chicago: University of Chicago Press, 1957.
- Bewley, Truman. "Why Not Listen to Business? The Study of Wage Rigidity." Unpublished manuscript, Yale University, 1997.
- Blinder, Alan S., et al. *Asking about prices: A new approach to understanding price stickiness*. New York: Russell Sage, 1998.
- Bobo, Lawrence D. "The Color Line, the Dilemma, and the Dream: Race Relations in America at the Close of the Twentieth Century," in John Higham, ed., *Civil rights and social wrongs: Black-white relations since World War II*. University Park: Pennsylvania State University Press, 1997, pp. 31-55.
- Brown, Claude. *Manchild in the promised land*. New York: Signet, 1965.
- Cogan, John. "The Decline in Black Teenage Employment." *American Economic Review*, September 1982, 72(4), pp. 621-38.
- Cose, Ellis. *The rage of a privileged class*. New York: Harper, 1993.
- Drake, St. Francis and Cayton, Horace. *Black metropolis*. Chicago: University of Chicago Press, 1945.
- Freeman, Richard B. and Holzer, Harry J., eds. *The black youth employment problem*. Chicago: University of Chicago Press, 1986.

- Geertz, Clifford. *The interpretation of cultures*. New York: Basic Books, 1973.
- Haley, Alex. *The autobiography of Malcolm X*. New York: Ballantine, 1965.
- Irons, Edward and Moore, Gilbert. *Black managers: The case of the banking industry*. New York: Praeger, 1985.
- Jaynes, Gerald David. *Branches without roots: Genesis of the black working class in the American South, 1862-1882*. New York: Oxford University Press, 1986.
- Johnson, Charles S. *Patterns of Negro segregation*. New York: Harper, 1943.
- Kirschenman, Joleen and Neckerman, Kathryn M. "We'd Love to Hire Them, But ...": The Meaning of Race for Employers," in Christopher Jencks and Paul E. Peterson, eds., *The urban underclass*. Washington, DC: Brookings Institution, 1991, pp. 203-32.
- Lewis, David Levering. *When Harlem was in vogue*. New York: Oxford University Press, 1981.
- McCall, Nathan. *Makes me wanna holler*. New York: Random House, 1994.
- Myrdal, Gunnar. *An American dilemma: The Negro problem and modern democracy*. New York: Pantheon, 1962.
- Ransom, Roger L. and Sutch, Richard. *One kind of freedom: The economic consequences of emancipation*. Cambridge: Cambridge University Press, 1977.
- Rees, Albert. "An Essay on Youth Joblessness." *Journal of Economic Literature*, June 1986, 24(2), pp. 613-28.

Quit Behavior as a Measure of Worker Opportunity: Black Workers in the Interwar Industrial North

By WARREN C. WHATLEY AND STAN SEDO*

Since Gary S. Becker (1957), economists have developed a variety of models to describe how the racial prejudices of workers, employers, and consumers lead to labor-market discrimination by race. Notable examples include Kenneth J. Arrow (1972, 1973), Edmund S. Phelps (1972), Dennis J. Aigner and Glen C. Cain (1977), Shelly J. Lundberg and Richard Startz (1983), Paul Milgrom and Sharon Oster (1987), and Dan A. Black (1995). As varied as these models are, each focuses on the wage component of the employment package, attempting to show how racial prejudices result in lower wages for black workers. Racial discrimination, however, can take a variety of nonwage forms. Racial differences in unemployment rates are well known and date back to at least 1930 (Ricard Vedder and Lowell Galloway, 1992). Robert E. B. Lucas (1974) finds that blacks have, on average, poorer nonwage job attributes than whites, and historically blacks have been overrepresented in seasonal, hot, dirty, and dangerous jobs (Sterling D. Spero and Abram L. Harris, 1931; Gunnar Myrdal, 1944). Of the major employers of black workers before World War II, the Ford Motor Company stands out for the opportunities it offered blacks; yet even Ford concentrated black workers in the foundry without paying them the compensating wage differential it had to pay white foundry workers (Christopher L. Foote et al., 1997). These results indicate that studies of racial discrimination that focus solely on wages are incomplete.

In this paper we investigate the structure of racial discrimination in U.S. labor markets between 1919 and 1943 in a way that incorporates information on the entire wage package. We do this by focusing on worker quit behavior rather than wages. The simplest of eco-

nomie assumptions says that a worker is more likely to quit his or her current job when the expected utility of an alternative employment package exceeds the expected utility of the current job. Black (1995) shows that, if some employers in the labor market discriminate against blacks in hiring, then black workers must search employers for a *racial* match in addition to the typical job match that other workers seek. In equilibrium, the additional monopsony power that employers have over black workers results in poorer job matches for black workers and lower reservation utilities. A lower reservation utility reduces job search and the propensity to quit. Also, since some employers do not hire black workers, these workers must search more firms to find a job. This increases the expected cost of job search, which further reduces the propensity to quit.

In this paper we use the personnel records of black and white workers hired between 1919 and 1943 to compare their conditional quit hazards. During this period, black workers became a significant presence in Northern labor markets for the first time. Nativistic and racialistic thinking was widespread, and racial conflict in the workplace periodically led to hate strikes and racially motivated social disturbances. Many Northern employers believed that black workers were inferior to whites, that they were best-suited for hot jobs, and that they were not intelligent or disciplined enough to keep up with the pace of work in Northern factories. Many white workers believed that black workers were different in another sense, that of being of a lower "civilization," and feared that their "willingness" to live on less would "tear down what the white man had built up" (Dwight Farnham, 1918; Chicago Commission on Race Relations, 1922; Spero and Harris, 1931; Charles S. Johnson, 1930; Myrdal, 1944; Raymond Wolters, 1970).

On the other hand, black workers complained of various kinds of discrimination, especially discrimination in employment offers,

* Department of Economics, University of Michigan, Ann Arbor, MI 48109.

job assignments, promotion opportunities, and treatment on the job. Their movement from job to job was often interpreted by employers as weak commitment to industrial work and a reason for the treatments about which they complained (Johnson, 1930; Spero and Harris, 1931; Myrdal, 1944; William H. Harris, 1982). This is what Myrdal called a "vicious circle," where initial white prejudices against blacks produced real racial differences in social and economic outcomes that in turn reinforced the belief among whites that blacks were inferior. A study of quit behavior can help us understand the logic of this process.

The samples we use are of male black and white workers hired at three large employers of blacks during this period: the Ford Motor Company plants of the Detroit area, the Pullman Standard Car Company in Chicago, and the Byers Steel Company in Pittsburgh. These samples are housed at the Inter-university Consortium for Political and Social Research (ICPSR) in Ann Arbor, Michigan, and are completely described in the code books stored with the data. They contain the complete work histories of workers employed at these plants, their occupations, length of time on the job, reasons for leaving, and individual worker characteristics.

The results of the analysis shed considerable light on the structure of labor-market discrimination in interwar America. We find that, on the surface, black workers may have appeared to be high-turnover workers, and anyone wanting such evidence could fashion it. A closer examination, however, shows that, once we control for occupations, black workers were much more stable than their white counterparts, suggesting that black workers' job opportunities were racially constrained. Married black men were the most stable of all workers, confirming an earlier finding by Thomas N. Maloney and Whatley (1995) that married black men in particular endured a difficult employment situation in the interwar North in their efforts to earn enough income to support their families. These results also show how racial prejudices in the North generated racial differences in economic behavior that many Northerners used to support their prejudicial priors, making it more difficult for Northern labor markets to reward the true productivity of black workers.

I. The Econometric Model

We measure racial differences in quit behavior by estimating a variant of the Cox proportional-hazards model proposed by Guido Imbens (1994). The Cox specification assumes stationarity in order for the parameter estimates to be consistent. In addition, all covariate effects that are specific to the spell must be parameterized. Since spells that are measured in duration time occur at different calendar times, this means that covariates such as spell-specific macroeconomic effects must be specified completely. These effects are often complicated combinations of seasonal and cyclical factors that may be difficult to specify correctly. Imbens (1994) proposes a variant of the Cox regression model that conditions on the calendar time of exit rather than spell duration. Using this specification, it is not necessary to parameterize the effects of calendar-time-specific covariates such as the unemployment rate. However, the effects of spell duration must now be specified within the hazard function.

This technique is particularly well suited to our personnel data. During the period in question, a number of economic events occurred which would certainly lend doubt to the assumption of stationarity in the macroeconomic environment. The boom of the 1920's was particularly strong in the automobile industry, while the Great Depression created a massive reduction in employment in all industries. A more complete discussion of this econometric technique can be found in Whatley and Sedo (1997a, b).

Although we use flow samples that have no censoring on duration, there is the possibility of multiple destinations. Information is available on whether a worker quit, was laid off, fired, or drafted into the military. Since it might be assumed that these would be the results of different characteristics, a competing-risks model is estimated. While the technique assumes independence between destinations, it does allow for estimation using the standard Cox regression model and therefore can be included in the method described by Imbens.

II. Empirical Results

The results from the proportional-hazard model with occupation controls are reported for the three firms in Table 1. We control for each occupation with at least 1 percent of the sample, which covers approximately 50 percent of the workers in each sample. When the model is estimated without these controls, the quit rates for black workers are typically much higher. For Ford, the estimate of the race coefficient without occupation controls is -0.174 , (meaning that blacks were less likely than whites to quit their jobs) and falls to -0.956 when occupation controls are added. For Byers, this coefficient falls from 0.215 to -0.056 , and for Pullman it falls from 1.508 to 1.422 . These changes indicate that black workers may have been disproportionately represented in high-turnover occupations, which means that some of the turnover of black workers can be attributed to characteristics of their jobs rather than the individuals themselves.

In addition to race, marital status deserves particular note. At Ford, both the black and black married coefficients are large, and the black coefficient is statistically significant. At Byers, the race coefficient is insignificant, but the marriage effect for black workers is extremely large and significant. Taken together, these results indicate that black workers at Ford and Byers have lower quit rates than their white counterparts.

The results for Pullman are different. Both the race and marriage effects are positive, and the race coefficient is statistically significant. However, these are offset by a large negative age effect that is not present at either Ford or Byers. The influence of this age effect is illustrated in Figure 1, which reports the relative quit rates for married black and white workers of various ages at the three firms. A value of 1 for this ratio indicates that married black workers quit at the same rate as their white counterparts, while values less than 1 indicate a lower quit rate for black workers. The middle line represents the point estimate of the relative quit rates, while the outer lines are the 95-percent confidence intervals for these estimates. As described above, married black workers at Ford and Byers are much less likely

TABLE 1—PROPORTIONAL-HAZARD MODEL

Independent variable	Coefficient	Standard error
<i>Ford</i> ($N = 2,521$):		
Duration	-12.785	0.7979
Duration \times duration	15.713	1.4506
Married	0.1633	0.2396
Hire age	-0.0147	0.0063
Dependents	-0.0129	0.0242
Previous employment	0.0429	0.1061
Married \times age	-0.0002	0.0083
Wage type	hourly	hourly
Black	-0.9563	0.4593
Black \times married	-0.6047	0.5013
Black \times age	0.0002	0.0152
Black \times dependents	0.0584	0.0444
Black \times previous employment	0.1043	0.1782
Black \times married \times age	0.0064	0.0179
Black \times wage type	hourly	hourly
<i>Byers</i> ($N = 5,031$):		
Duration	-21.894	0.9114
Duration \times duration	19.101	1.0383
Married	0.6888	0.1636
Hire age	-0.0078	0.0036
Dependents	-0.0345	0.0178
Previous employment	-0.5933	0.0564
Married \times age	-0.0158	0.0048
Wage type	0.7137	0.1683
Black	-0.0564	0.1893
Black \times married	-0.7877	0.2674
Black \times age	-0.0062	0.0066
Black \times dependents	0.0045	0.0309
Black \times previous employment	0.0117	0.0922
Black \times married \times age	0.0189	0.0085
Black \times wage type	hourly	hourly
<i>Pullman</i> ($N = 1,369$):		
Duration	-47.824	2.4145
Duration \times duration	67.482	6.1663
Married	-0.1857	0.3866
Hire age	0.0085	0.0098
Dependents	-0.0769	0.0474
Previous employment	-0.4185	0.1389
Married \times age	0.0025	0.0121
Wage type	-0.2166	0.1543
Black	1.4217	0.4507
Black \times married	0.1693	0.5794
Black \times age	-0.0462	0.0148
Black \times dependents	0.0759	0.0643
Black \times previous employment	-0.0478	0.1845
Black \times married \times age	-0.0071	0.0189
Black \times wage type	0.0421	0.1866

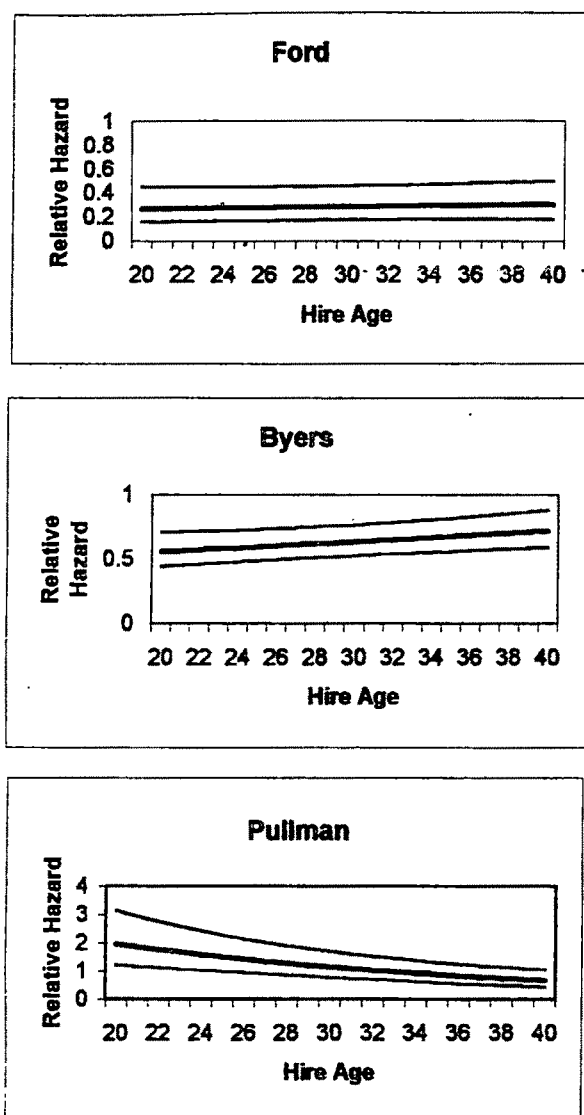


FIGURE 1. RELATIVE QUIT RATES FOR BLACK AND WHITE WORKERS

to quit than are married white workers, regardless of age. While in both cases the relative quit rate for black workers increases with age, the upper bound for the 95-percent confidence interval is always less than 1 for all age categories.

The results for Pullman show that, while young black workers have a higher quit rate than young white workers, this effect diminishes with age. By age 25, the lower bound of the 95-percent confidence interval is less than 1, and by age 33 the point estimate is less than that of white workers. To put these results in perspective, it is useful to consider the typical married black worker in each of the three

firms. At Ford and Byers, the median hire age is 29. The point estimate for a 29-year-old married black worker is 0.283 for Ford and 0.723 at Byers, and the upper bound of the 95-percent confidence interval is less than 1 in both cases, as described above. For Pullman, the median hire age is 30. The point estimate for a 30-year-old married black worker is 1.147, which indicates a slightly higher quit rate. However, the lower bound of the 95-percent confidence interval is 0.778, which means that this difference is not statistically significant. These results show that, relative to white workers, the typical married black worker is less likely to quit at Ford and Byers and no more likely to quit at Pullman, which contradicts the notion that black workers were not as stable as whites.

III. Conclusion

The results described above show that black workers had quit rates in the interwar period that were typically lower than those of their white counterparts. For Ford and Byers, these lower quit rates were present regardless of the age at which the workers were hired. In addition, the evidence indicates that some of the quit behavior attributed to black workers may in fact be the result of their overrepresentation in jobs that have high turnover rates for all types of workers.

These results also help us to understand the dynamics of vicious circles. Racial prejudice forced black workers to value their jobs more highly than whites, but the jobs they received were in high-turnover occupations. Those wanting to find evidence of racial differences to support racialistic priors did not have to look hard, so long as they did not look too closely. Seeing the reality required a more critical stance, one not easily taken then or now.

REFERENCES

- Aigner, Dennis J. and Cain, Glen C. "Statistical Theories of Discrimination in Labor Markets." *Industrial and Labor Relations Review*, January 1977, 30(1), pp. 175-87.
- Arrow, Kenneth J. "Some Mathematical Theories of Race in the Labor Market," in Anthony Pascal, ed., *Racial discrimination*

- in *economic life*. Lexington, MA: Heath, 1972.
- . "The Theory of Discrimination," in Orley Ashenfelter and Albert Rees, eds., *Discrimination in labor markets*. Princeton, NJ: Princeton University Press, 1973.
- Becker, Gary S. *The economics of discrimination*. Chicago: University of Chicago Press, 1957.
- Black, Dan A. "Discrimination in an Equilibrium Search Model." *Journal of Labor Economics*, April 1995, 13(2), pp. 309–34.
- Chicago Commission on Race Relations. *The Negro in Chicago*. Chicago: Chicago University Press, 1922.
- Farnham, Dwight. "Negroes as a Source of Industrial Labor." *Labor Management*, August 1918, 56, pp. 123–29.
- Foote, Christopher L.; Whatley, Warren C. and Wright, Gavin. "Arbitraging a Discriminatory Labor Market: Black Workers and the Ford Motor Company, 1918–1947." Mimeo, University of Michigan, 1997.
- Harris, William H. *The harder we run*. New York: Oxford University Press, 1982.
- Imbens, Guido. "Transition Models in a Non-stationary Environment." *Review of Economics and Statistics*, September 1994, 76(3), pp. 703–20.
- Johnson, Charles S. *The Negro in American civilization*. New York: Holt, 1930.
- Lucas, Robert E. B. "The Distribution of Job Characteristics." *Review of Economics and Statistics*, November 1974, 56(4), pp. 530–40.
- Lundberg, Shelly J. and Startz, Richard. "Private Discrimination and Social Intervention in Competitive Labor Markets." *American Economic Review*, June 1983, 73(3), pp. 340–47.
- Maloney, Thomas N. and Whatley, Warren C. "Making the Effort: The Contours of Racial Discrimination in Detroit's Labor Market, 1920–1940." *Journal of Economic History*, September 1995, 55(3), pp. 465–93.
- Milgrom, Paul and Olster, Sharon. "Job Discrimination, Market Forces, and the Invisibility Hypothesis." *Quarterly Journal of Economics*, August 1987, 102(3), pp. 453–76.
- Myrdal, Gunnar. *An American dilemma*. New York: Harper, 1944.
- Phelps, Edmund S. "The Statistical Theory of Racism and Sexism." *American Economic Review*, September 1972, 62(3), pp. 659–61.
- Spero, Sterling D. and Harris, Abram L. *The black worker*. New York: Columbia University Press, 1931.
- Vedder, Richard and Galloway, Lowell. "Racial Differences in Unemployment in the United States, 1890–1990." *Journal of Economic History*, September 1992, 42(3), pp. 692–702.
- Whatley, Warren C. and Sedo, Stan. "The Historical Origins of Internal Labor Markets: Evidence from Personnel Records." Mimeo, University of Michigan, 1997a.
- . "The Unreliability of Job Duration Estimates, Past and Present." Mimeo, University of Michigan, 1997b.
- Wolters, Raymond. *Negroes and the Great Depression*. Westport, CT: Greenwood, 1970.

Assessing 50 Years of African-American Economic Status, 1940–1990

By MARCUS ALEXIS*

On the eve of World War II, African-Americans suffered the twin afflictions of (i) segregated education, housing, transportation, and public accommodations and (ii) employment discrimination, which translated into lower earnings, entry barriers to many occupations and industries, and greater exposure to the business cycle and to competition at home and overseas. Post-World War II labor-market conditions (strong aggregate demand and relatively improved labor quality of African-American workers, civil rights legislation, and supportive court decisions) reduced job barriers and narrowed pay differentials. As a consequence, relative earnings of African-Americans increased in the aggregate and in subgroups from 1940 to 1970. The African-American/white earnings gap narrowed until the late 1970's and has since stagnated, an effect which was not apparent in early comparisons of the 1970 and 1980 Census data.

I. Benchmarking Earnings Differentials

James P. Smith and Finis Welch (1989) computed earnings of African-American and white male workers over the 1940–1980 period. The comparison is weekly wages of full-time working males. An obvious shortcoming of this approach is that unemployed workers, part-time workers, and women are excluded. The results are important, nevertheless, in identifying changes in the relative hourly earnings of fully employed men between 1940 and 1980.

To remove the effects of inflation Smith and Welch express their data in constant 1987 dollars. The base year is 1940, and the endpoint is 1980. In the base year (1940), the ratio of the median weekly wage of African-American

men to that of white men was 43.3 percent, ranging from 39.8 percent for those with 36–40 years experience to 47.5 percent for those with 6–10 years experience. For the least experienced (1–5 years) it was 46.7 percent. Older African-American men had lower relative wages.

By 1950 the overall ratio of African-American to white weekly wages increased by 12 percentage points to 55.2 percent. Even the oldest (most experienced) men saw their relative wage rise several percentage points to 46.9 percent. The most impressive gains were among the youngest (1–5 years experience) men, rising from 46.7 percent to 61.8 percent, 15 percentage points. The large relative change reflected both the depressed base year (1940) and the strong post-World War II economy. The job market explains the general rise in the wages of both African-American and white men, but not the relative change. Two factors were likely to have been at work: a tighter labor market presented opportunities for African-American men to move up the occupational ladder, and market discrimination eased due to both the relative abundance of jobs and the lessening of overt discrimination.

During the next decade (1950–1960), African-American men experienced virtually no increase in overall relative wages, a gain of 2 percentage points, from 55.2 percent to 57.5 percent. The entire gain is explained by the exit of two cohorts, those entering the labor force before 1923. Younger cohorts with less than ten years experience actually did slightly worse than similarly experienced workers in 1950.

Between 1960 and 1970 a combination of older workers with lower relative wage exiting and a 15-percentage-point gain by younger workers (those with less than five years experience) and modest improvements at all other experience levels pushed the 1970 overall relative wage to 64.4 percent, up 6.9 percentage points over the 1960 level.

* Kellogg Graduate School of Management, Northwestern University, 2001 Sheridan Road, Evanston, IL 60208.

In 1980 the overall relative percentage increased to 72.6, up 8.2 percentage points, and virtually all experienced cohorts experienced some gain. But again, it was the younger cohorts that did best. The relative wage increased to 76.6 percent for African-American men with 5–10 years experience, and those with less than five years experience received 84.2 percent of the wage received by their white counterparts.

Over the 40-year period, full-time employed African-American men with less than five years of experience moved from 46.7 percent of the white wage to 84.2 percent. Those with 36–40 years experience had their relative wage increase from 39.8 percent to 68.5 percent, impressive gains. Even if one takes the more representative post-World War II base year of 1950, the gains for the least experienced African-Americans went from 61.8 percent to 84.2 percent, up 22.4 percentage points. And for the most experienced men, it increased from 46.9 percent to 68.5 percent, a gain of 21.6 percentage points, not quite the values of 37.5 percent and 28.7 percent, respectively, of the 1980–1940 comparison, but quite respectable. Had the narrowing rate of 1950–1980 continued, African-American men would have achieved parity by 2020.

II. Labor-Force Participation

Labor-force participation rates between 1940 and 1970 for both African-American and white men were relatively stable. After 1970 both races experienced sharp declines in participation rates. The decline was much larger for African-American men, most notably for those of prime working age. The decline for African-American men 36–45 years old was 5–6 percentage points versus 1.3 percentage points for white men in the same age group. During the decade of the 1970's, at ages 46–54, labor-force participation of African-American men fell 9.8 percentage points; for white men of the same ages, labor-force participation declined a much smaller 3.5 percentage points. In both cases exit rates were highest among less well-educated men, highest for those with 0–7 years of schooling, followed by those with 8–11 years.

III. Unemployment

In the early 1970's the ratio of African-American to white unemployment (male and female) hovered around 2. Following a brief jump in the late 1970's to nearly 2.5, the ratio fell steadily to 2.2 in 1982. During the 1980's there was a steady progression, so that by the end of the decade the ratio was slightly in excess of 2.5. Unemployment rates of African-Americans during the economic expansion of the 1980's fell from 18.9 percent in 1982 to 11.3 percent in 1989; for whites it fell from 8.6 percent to 4.5 percent.

For younger African-Americans, those 16–19 years old, the relative-unemployment ratio is virtually constant from the early 1970's (1972) to 1989, roughly 2.5. One interpretation of this is that unskilled workers of both races and sexes experienced the same market forces. An alternate explanation is that younger African-Americans were more likely to have exited the labor market, thus understating the relative joblessness of young African Americans.¹

An interesting insight into the labor-market difficulties facing young African-Americans is reported by Kathryn M. Neckerman and Joleen Kirschman (1990). They found that employers view inner-city African-American youth as having poor skills, bad work habits, and no work ethic. This view of inner-city African-American youths did not extend to suburban African-American youths. Applicants' addresses were used as a screen. Addresses in large public-housing projects were a potent negative signal. Inner-city public schools were considered a poor place to recruit. School recruiting of African-American youth was mainly from private and parochial schools. Employee reference networks were important sources of new workers, and recruitment was centered on particular neighborhoods. Job interviews were also given great weight. Similar results were reported by

¹ Lisa M. Lynch (1989) finds that the longer both young men and women are unemployed, the less likely they are to find another job. Her data show that this reduced probability of reemployment is greater for nonwhites.

Jomills Henry Braddock and James M. McPartland (1990). An important difference is that they found some African-Americans to have adequate employment networks, but those from segregated high schools with segregated networks disproportionately held jobs paying lower wages; this effect is much stronger in entry-level jobs than in those that are the results of a promotion. These results are consistent with those found by Richard Price and Edwin Mills (1985). Using alternate specifications to measure the effects of race and suburban versus central-city location they estimate that central-city location lowers earnings by between 5.7 percent and 8.6 percent.

Using U.S. Census data, Glenn G. Cain and Ross E. Finnie (1980) studied labor-market conditions for African-American youth in 94 statistical metropolitan areas. They found a modest but statistically significant relationship between the African-American male unemployment rate and the number of hours worked. A 1-percent decline in the unemployment rate increased hours worked annually by 2 percent. A much stronger relationship between local labor-market conditions and employment outcomes of young African-American men is found by Richard B. Freeman (1991). For all youth, a 1-percent decline in unemployment raises the employment/population ratio 4.3 percentage points. A consistent finding in both Cain and Finnie (1980) and in Freeman (1991) is that the relative position of African-American youth is very sensitive to local labor-market demand for all labor.

Joblessness among young African-American men has been of concern to policymakers throughout the 1970's and 1980's. Reasons for this high rate of joblessness include the facts that: (i) demand for less-skilled labor has decreased; and (ii) the impact of this decline falls disproportionately on African-American men. The clearest indication that industrial shifts have worsened the relative employment position of African-American men is from John Bound and Harry J. Holzer (1990). Their cross-sectional study of 52 metropolitan areas finds that manufacturing shifts account for as much as 10–30 percent of the employment declines for subgroups of African-American and white men and are larger for the former group. Between 35 per-

cent and 50 percent of the employment decline was experienced by African-American dropouts aged 16–24 between 1970 and 1980. Based on these results, one would have predicted further declines in the employment in this age group during the 1980's as the manufacturing share of gross domestic product continued to fall, a prediction which turned out to be correct. Support for Bound and Holzer's results is provided by Barry Bluestone et al. (1991), who report that one-third of the increase in year-round unemployment of 20-year-old, less-educated, young African-American males and 100 percent of the decrease in the weeks worked are due to decreased manufacturing employment shares.

A somewhat different result is reported by Welch (1990), for the decade of the 1970's. He constructs an index of industry shifts from 1940 to 1980 using Census data. His units of observation are states, not metropolitan areas, and he includes all major industrial shifts, not just manufacturing. While he does not find a strong link between industrial shifts and reduced demand for African-Americans in the 1970's, he does find that agricultural shifts reduced labor demand from 1940 to 1960. But the magnitude of such shifts in the 1970's is much smaller. When Welch attempts to apply his methodology to metropolitan areas, estimated effects are unstable. Smith and Welch (1989 pp.550–51) attribute the observed results to African-American youth labor-supply conditions:

[T]he main culprit is that the reservation wage of black youth (the minimum wage they are willing to accept) has increased faster than the rising market wages of black youth.

This view is challenged by Holzer (1986) based on data from the National Longitudinal Survey of Youth and the National Bureau of Economic Research 1980 survey of inner-city youths in Boston, Chicago, and Philadelphia. He finds the reservation wages of African-American youth to be quite similar to that of white youth. Observed wages for whites are, however, higher. African-American youth are more likely to accept low paying jobs than whites. Alexis et al. (1983a, b) found similar results in a comparison of Latinos, African-

Americans, and whites in the Chicago labor market.

Wage aspirations of inner-city African-American youths were higher than actual wages in the 1970's. Thus, these youths were more likely to view their jobs as temporary and to be less attached to them, giving rise to longer spells of joblessness than is observed for white youths. It can be argued that similar reservation wages mask dissimilar human-capital endowments and higher effective African-American reservation wages. Relatively greater improvements in labor-market skills over time by African-American men, as posited by Smith and Welch (1989) reduces the likelihood of this effect. Taken as a whole, the evidence does not support reservation-wage disparities as a contributing factor to a worsening job market for African-American men.

Richard Butler and James J. Heckman (1977) speculated that the support provided by welfare acted as a reservation-wage floor and that higher levels of unemployment could be explained by this relationship. Whatever the relationship between welfare and joblessness, there is no evidence to support an increasing attractiveness of welfare in the 1970's as a cause for worsening joblessness rates for African-American males. In the 1980's the real value of welfare declined. This decline in the real value of welfare did not reverse the trend of greater African-American joblessness which may be traced back to the 1970's. There is evidence, however, that income-earning opportunities from crime grew in the 1980's. The growth in the market for drugs appears to have been a major outlet for illegal activities and associated income.

During the 1970's the apparent tendency toward convergence of market outcomes for African-Americans and whites ceased, and a reverse movement that erased some of the post-World War II gains ensued. Much of the focus of research in the 1980's was on younger men's outcomes. This focus may prove to be helpful if one wishes to project whether incomes (wages) of African-American men will again experience a wave of convergence or whether the relative status will stagnate or diverge. First, young African-American men are more like their counterparts in education than

are older racial cohorts. Second, the experiences, skills, and attachment to work of this group of young men will affect their labor-market outcomes in later years. Third, their labor-market experiences may feed back to affect later generations of young African-Americans through the networks formed or missed and through the family and neighborhood (community) interactions. The latter are very important and could alter, for instance, (i) decisions on labor-force participation, (ii) length and type of schooling, (iii) the choice of private-market or military occupations, (iv) allocations of time between "straight" and "extra-market" (underground and criminal) activities. A literature on these dynamic interactions is accumulating but is beyond the scope of this paper.²

Younger, less-educated workers are most affected by adverse labor-market conditions (Cain and Finnie, 1990; Freeman, 1991). Changes in the composition of industry were responsible for up to one-third the increase in year-round joblessness among less-educated African-American men (Bluestone et al., 1991). African-American men experienced greater difficulty shifting from high-paying manufacturing jobs to equally remunerative nonmanufacturing jobs in the growing service sector; they were also less successful in finding jobs in the higher-paying professional, technical, managerial, and sales sectors (David Howell, 1991). The impact of this as noted by Jeremiah Cotton (1989) was a decline in the relative earnings of African-American men to 0.79 in blue-collar jobs and to 0.67 in white-collar jobs.

The shifts of African-American men out of manufacturing is most pronounced in durable-goods manufacturing. Between 1973 and 1975, 20.1 percent of African-American male workers and 18.4 percent of white male workers worked in durable-goods manufacturing. During 1987-1988, the percentages were 10.2 and 14.4, respectively, nearly a 50-percent decline for African-Americans (49.2 percent)

² Representative papers are Robert D. Mare and Christopher Winship (1984), William J. Wilson and Neckerman (1986), Anne Case and Lawrence F. Katz (1990), and Freeman (1991).

versus a 21.7-percent decline for whites. For African-American men, the Midwest was an unmitigated disaster as far as durable-goods manufacturing was concerned. During 1973–1974, 42.1 percent of Midwestern African-American men and 33.0 percent of white men were employed in durable-goods manufacturing. By 1987–1988 the respective percentages were 12.5 and 21.3. The decline for African-Americans was 70.3 percent; for whites, 35.4 percent. In the United States overall and in the Midwest, both races suffered large job losses, but white losses were only one-half those of African-Americans.

IV. Immigration

Conventional wisdom has suggested that the large waves of immigration in 1970's and 1980's reduced job opportunities for low-skill workers. Identifying such an effect has proved to be elusive. Cross-sectional studies (Alexis and Nancy DiTomaso, 1983; Alexis et al., 1983a, b; George J. Borjas, 1983, 1987; Joseph G. Altonji and David Card, 1991) have failed to find large consistent effects.

Borjas et al. (1991) use the 1980 Census and Current Population Survey (CPS) data (through 1987) to estimate the effect(s) of immigration on less-educated youths and find surprisingly large impacts. In addition to the Census and CPS data, they use estimates of implicit increases in the supply of less-educated labor embodied in net imports of the 1980's. They find that up to one-third of the observed change in the differential between college- and high-school-level labor may be attributed to the labor-supply effects of immigration and trade.

The findings of Borjas et al. (1991) represent significant advances over earlier studies for several reasons: (i) earlier studies use data from 1980 or earlier; (ii) the earlier studies miss most recent periods, when employment and earnings problems of African-American men and problems of concentrated poverty appeared to have worsened; (iii) earlier studies concentrate primarily on earnings and unemployment of working populations, and they exclude the effects of immigration on the nonworking jobless; (iv) studies using national cross sections miss the interaction of im-

migration flows and growth of disadvantaged populations which appear to be concentrated in relatively few cities; (v) previous studies looked at either immigration or trade but not at both together in an integrated cohesive way.

V. Discrimination

An audit study of employer discrimination by Raymond J. Struyk and his colleagues at the Urban Institute is a reminder that discrimination in all its forms—overt and statistical—still exists (see Struyk et al., 1991). The Urban Institute team sent African-Americans and whites with matched sets of educational backgrounds, experiences, and demeanors to apply for the same jobs in Washington, DC, and Chicago. The results were that white men were three times as likely as African-American men (20 percent vs. 7 percent) to advance in the hiring process; offers were received by the white applicant 15 percent of the time and by the African-American applicant 5 percent of the time.

The Neckerman and Kirschenman (1990) results reported earlier also point to continued biased perceptions by employers (a mixture of race and class bias). Braddock and McPartland (1990) find that employers are less likely to hire minorities for jobs that emphasize academic achievement and cerebral activity. Finally, indicators of Equal Employment Opportunity Commission enforcement activity in the 1980's are consistent with reduced anti-discrimination enforcement: new cases (fewer), class-action suits (fewer), and staff size (down 20 percent).

VI. African-American Women

Francine M. Blau and Arlene H. Beller (1991) use the 1972, 1982, and 1989 Annual Demographic Files of the CPS to study the effect of part-time employment and annual weeks worked on race and gender earnings differentials. Cutoff weeks worked and annual earnings are respectively one week and \$100, permitting observation over a wide range of experiences. They find that relative racial earnings stagnated during the 1980's for both men and women, but the pattern is

a reverse of the earlier years (1960's). For both male and female African-Americans, younger cohorts did not do better than older, more experienced ones. The largest gains over the 1971-1988 groups was found in African-American females with 20+ years of experience. These women narrowed the earnings gap relative to all groups: white women, white men, and African-American men. Likewise older African-American men were the only ones to improve their relative position during the period. These observations fly in the face of predictions of racial earnings convergence based on greater relative improvement in African-American workers. A possible explanation is that African-American inputs are no longer improving faster than white inputs. Another possibility is that there are structural changes in the labor market which disproportionately hurt younger African-Americans because (i) they are less experienced, (ii) they have the wrong skills (poor counseling/curriculum), (iii) they face greater competition from imports, immigrants, and females entering the labor market, and (iv) they possess inadequate job search skills.

The complexity of labor-market outcomes is reflected in the Blau and Beller (1991) finding that African-American females who entered the labor market in the 1960's not only retained, but increased their relative wages as they aged. Women who entered in the 1970's had a lower relative wage than the women who entered a decade earlier.

Another difference in the 1970's experience of men and women involved occupation. For women with 20+ years of experience, the effect was positive and substantial, largely reflecting an exodus from private household (domestic) work. During the 1980's, occupational shifts played a much smaller role in reducing wage inequality. For men, occupational shifts worked in the other direction, increasing earnings differentials, because of the underrepresentation of African-American men in managerial and professional jobs.

Mary Corcoran and Sharon Parrott (1997) reach many of the same conclusions as Blau and Beller (1991) with respect to the stagnation of African-American earnings relative to those of comparable white men and women in

the 1980's and the relatively superior labor-market experience of older women. But Corcoran and Parrott concentrate on younger African-American women whose employment and earnings they believe better reflect changing economic conditions than do the labor-market outcomes of older African-American women.

Among college graduates Corcoran and Parrott find steady decline in the relative earnings of African-American women in the 1970's and 1980's. For other schooling groups they find a decline in the earnings gap in the 1980's following the stable 1970's. The decline is statistically significant for high-school graduates but not for the other groups. They also find that earnings trends differ by education group and region. For reasons not explained, the racial gaps in earnings of both male and female college graduates widen considerably in both the 1970's and 1980's. At the other end of the educational spectrum is the high-school dropout. Both high-school dropouts and single mothers have experienced large declines in employment since 1969. Throughout the 1970's and 1980's the gap between young white and young African-American women's employment opportunities widened. The rising labor-force participation rate of white college graduates in the 1970's and 1980's reduced the cost to employers who favored whites.

The lesson of both Blau and Beller (1991) and Corcoran and Parrott (1997) is that the labor-market experiences of African-American men and women can be and often are different. One needs to consider gender as well as race to fully understand African-American labor-market outcomes.

VII. Conclusions

I have examined the postwar economic experience of African-American men and observed a pattern of relative improvement in weekly earnings between the census years of 1940 and 1980. The same pattern holds if the base year is 1950. Improvements in years of schooling and schooling quality, South-North migration, and reduced discrimination in the South are all credited with the narrowing of the racial gap in earnings.

In the mid-1970's, the narrowing of racial disparities stagnated, and it reversed in the 1980's. Many factors are associated with the 1980's decline in the relative positions of African-American men: (i) a shift in labor demand toward more skilled jobs; (ii) changing industrial and occupational distribution by race, which accounts for one-third of the widening differential; (iii) a disproportionately negative impact of diminishing blue-collar employment opportunities; (iv) the disproportionately negative impact on central cities of the decline in blue-collar employment; (v) overrepresentation of African-Americans in hard-hit regions (e.g., the Midwest); (vi) the location of African-American residences in the wrong areas within the urban areas; (vii) competition from immigrants and imported goods; (viii) the continuing presence of overt and statistical discrimination; (ix) diminished government enforcement of antidiscrimination laws; and (x) physical and psychological isolation.

Studies of labor-market outcomes of African-American women also reveal earnings gaps in the 1960's, 1970's, and 1980's, but the pattern differs somewhat from the male experience. The large relative gains of younger male cohorts narrowed in the 1970's and stagnated in the 1980's while older African-American women experienced a continued reduction in cohort earnings differentials. Younger, particularly less-well-educated African-American women did not progress as much as those with 20+ years of experience. Dropouts and single mothers experienced declines in relative position.

REFERENCES

- Alexis, Marcus and DiTomaso, Nancy. "Income Determination in Three Internal Labor Markets." Mimeo, Northwestern University, 1983.
- Alexis, Marcus; DiTomaso, Nancy and Kyle, Charles. "Hispanic Adaption in the Chicago Labor Market," in *Proceedings of the Rockefeller Foundation Workshop on the Labor Market of Immigrants*. New York: Rockefeller Foundation, 1983a.
- . "Impact of Hispanic Immigration On Black Workers in Chicago." Mimeo, Center for Urban Affairs, Northwestern University, 1983b.
- Altonji, Joseph G. and Card, David E. "The Effects of Immigration on the Labor Market Outcomes of Less-Skilled Natives," in John M. Abowd and Richard B. Freeman, eds., *Immigration, trade, and the labor market*. Chicago: University of Chicago Press, 1991, pp. 201–34.
- Blau, Francine M. and Beller, Arlene H. "Black-White Earnings in the 1970s and 1980s: Gender Differences." National Bureau of Economic Research (Cambridge, MA) Working Paper No. 3736, 1991.
- Bluestone, Barry; Stevenson, Mary and Tilly, Chris. "The Deterioration in Labor Market Prospects for Young Men with Limited Schooling: Assessing the Impact of 'Demand Side' Factors." Unpublished manuscript presented at Eastern Economics Association Meeting, Pittsburgh, PA, 14–15 March 1991.
- Borjas, George J. "The Substitutability of Black, Hispanic, and White Labor." *Economic Inquiry*, January 1983, 21(1), pp. 93–106.
- . "Immigrants, Minorities, and Labor Market Competition." *Industrial and Labor Relations Review*, April 1987, 40(3), pp. 382–92.
- Borjas, George J.; Freeman, Richard and Katz, Lawrence P. "On the Labor Market Effects of Immigration and Trade." National Bureau of Economic Research (Cambridge, MA) Working Paper No. 3761, 1991.
- Bound, John and Holzer, Harry J. "Industrial Shifts, Skill Levels, and the Labor Market for White and Black Males." Mimeo, University of Michigan, 1990.
- Braddock, Jomills Henry, II and McPartland, James M. "How Minorities Continue To Be Excluded from Equal Employment Opportunities: Research on Labor Market and Industrial Barriers." *Journal of Social Sciences*, January 1990, 9, pp. S376–95.
- Butler, Richard and Heckman, James J. "The Government's Impact on the Labor Market, Status of Black Americans: A Critical Review," in Farrel E. Block et al., eds., *Equal rights and industrial relations*. Madison, WI: Industrial Relations Research Association, 1977, pp. 235–81.

- Cain, Glenn G. and Finnie, Ross E. "The Black-White Difference in Youth Employment: Evidence for Demand-Side Factors." *Journal of Labor Economics*, January 1990, 8(1), part 2, pp. S364-95.
- Case, Anne and Katz, Lawrence F. "The Company You Keep: The Effect of Family and Neighborhood on Disadvantaged Youths." Mimeo, Harvard University, 1990.
- Corcoran, Mary and Parott, Sharon. "Black Women's Economic Progress." Mimeo, University of Michigan, 1997.
- Cotton, Jeremiah. "Opening the Gap: The Decline in Black Economic Indicators in the 1980s." *Social Science Quarterly*, December 1989, 40(4), pp. 803-19.
- Freeman, Richard B. "Labour Market Tightness and the Mismatch Between Demand and Supply of Less-Educated Young Male Workers in the United States in the 1980s," in Fiorella Padoa-Schioppa, ed., *Mismatch and labour mobility*. Cambridge: Cambridge University Press, 1991, pp. 360-81.
- Holzer, Harry J. "Black Youth Non-employment: Duration and Job Search," in Richard B. Freeman and Harry Holzer, eds., *The black youth employment crisis*. Chicago: University of Chicago Press, 1986, pp. 23-70.
- Howell, David. "Economic Restructuring and Employment Status of Young Black Men: 1979-1989." Mimeo, Graduate School of Management, New School of Social Research, 1991.
- Lynch, Lisa M. "The Youth Labor Market in the Eighties: Determinants of Reemployment Probabilities for Young Men and Women." *Review of Economics and Statistics*, February 1989, 71(1), pp. 37-45.
- Mare, Robert D. and Winship, Christopher. "The Paradox of Lessening Racial Equality and Joblessness Among Black Youth: Enrollment, Enlistment, and Employment, 1964-1981." *American Sociological Review*, February 1984, 49(1), pp. 39-55.
- Neckerman, Kathryn M. and Kirschenman, Joleen. "Hiring Strategies, Racial Bias and Inner City Workers: An Investigation of Employers' Hiring Decision." Mimeo, Department of Sociology, University of Chicago, 1990.
- Price, Richard and Mills, Edwin. "Race and Residence in Earnings Determination." *Journal of Urban Economics*, January 1985, 17(1), pp. 1-18.
- Smith, James P. and Welch, Finis. "Black Economic Progress After Myrdal." *Journal of Economic Literature*, June 1989, 27(2), pp. 519-64.
- Struyk, Raymond J.; Turner, M. A. and Fix, Michael. *Opportunities denied, opportunities diminished: Discrimination in hiring*. Washington, DC: Urban Institute, 1991.
- Welch, Finis. "The Employment of Black Men." *Journal of Labor Economics*, January 1990, 8(1), part 2, pp. S26-74.
- Wilson, William J. and Neckerman, Kathryn. "Poverty and Family Structure: The Widening Gap Between Evidence and Public Policy Issues," in Sheldon Danziger and Daniel Weinberg, eds., *Fighting poverty: What works and what doesn't*. Cambridge, MA: Harvard University Press, 1986, pp. 232-59.

THEORETICAL AND EMPIRICAL DEVELOPMENTS IN COST-BENEFIT ANALYSIS AND PROGRAM EVALUATION[†]

Imagined Risks and Cost-Benefit Analysis

By ROBERT A. POLLAK*

Everyone recognizes substantial discrepancies between the public's rankings of hazards and those of the experts. For example, experts at the Environmental Protection Agency think that hazardous-waste sites pose "medium-to-low" risks to the public, while indoor air pollution poses a "high" risk; yet public perceptions have driven policy to focus on hazardous-waste sites rather than on indoor air quality (Stephen Breyer, 1993 pp. 19–20).

Whose beliefs should determine government policy when the public's beliefs differ from those of the experts? The problem evaporates if the public, perhaps recognizing its inability to deal with complex technical issues, entrusts risk assessment to the government and its experts. But what if the public, perhaps distrusting government and experts, is unwilling to leave risk assessment to the experts?¹ Paul Portney (1992 p. 131) posed a version of this "Whose beliefs?" question succinctly in his fable, "Trouble in Happyville":

You have a problem. You are Director of Environmental Protection in Happyville, a community of 1000 adults. The drinking water supply in Happyville is contaminated by a naturally occurring

substance that each and every resident believes may be responsible for the above-average cancer rate observed there. So concerned are they that they insist you put in place a very expensive treatment system to remove the contaminant. Moreover, you know for a fact that each and every resident is truly willing to pay \$1000 each year for the removal of the contaminant.

The problem is this. You have asked the top ten risk assessors in the world to test the contaminant for carcinogenicity. To a person, these risk assessors—including several who work for the activist group, Campaign Against Environmental Cancer—find that the substance tests negative for carcinogenicity, even at much higher doses than those received by the residents of Happyville. These ten risk assessors tell you that while one could never prove that the substance is harmless, they would each stake their professional reputations on its being so. You have repeatedly and skillfully communicated this to the Happyville citizenry, but because of a deep-seated skepticism of all government officials, they remain completely unconvinced and truly frightened—still willing, that is, to fork over \$1000 per person per year for water purification.

[†] *Discussants:* Sherwin Rosen, University of Chicago; W. Kip Viscusi, Harvard University. Paul R. Portney (Resources for the Future) also presented a paper in this session ("Cost-Benefit Analysis: The Need for Straight Talk") but elected not to publish it in this volume.

* Department of Economics and John M. Olin School of Business, Washington University, St. Louis, MO 63130. This article draws on Pollak (1995, 1996, 1998). I am grateful to Judith Goff for editorial assistance.

¹ I treat "the public" and "the experts" as if they were homogeneous groups. I shall not discuss the case in which the public is divided, the experts are divided, and claims of expertise are countered by accusations of "junk science."

Portney does not pose the mirror-image case, exemplified by indoor air pollution, in which the public is less concerned than the experts. Later in the article, however, he recasts his hypothetical case from the regulatory to the legal context, replacing the naturally occurring substance with an industrial contaminant. The toxic tort version of the question becomes: should those who believe they have been harmed receive compensation from those responsible for the contamination? A related policy question is posed by the concerns of

people living near an abandoned hazardous-waste site who worry that it will cause them or their children to develop cancer even though experts "know" that these fears are unwarranted.²

Most policy analysts would agree with Stephen Breyer, now an Associate Justice of the U.S. Supreme Court, that discrepancies between the public's rankings of hazards and those of the experts "reflect not different values but different understandings about the underlying risk-related facts" (Breyer 1993 p. 35). They would also agree with Breyer that government policy ought to be based on the risk-related facts and not on the public's (mis)perceptions. But Breyer assumes rather than argues this conclusion. Portney's *Happyville* fable, by posing the "Whose beliefs?" question succinctly, invites a discussion of the roles of public and expert perceptions in risk assessment and cost-benefit analysis. Thus far, few have accepted the invitation.

I. Why Perceptions Differ

Breyer (1993) offers a laundry list of explanations of why public perceptions differ from those of experts: people have difficulty with the mathematics of probability; they respond to the framing of issues; they use heuristics or rules of thumb that are misleading in situations involving low-probability events; they distrust experts; and they are exposed to sensational stories in the media. Howard Margolis (1996) argues convincingly that many of the standard explanations of divergence between public and expert perceptions are not explanations at all, but merely reassertions that public and expert perceptions diverge. For example, he points out that it is at least as plausible that the public distrusts the experts *because* it believes nuclear-waste disposal is unsafe as that it believes nuclear-waste disposal is unsafe *because* it distrusts the experts (Margolis, 1996 p. 29).

² Under Superfund's "polluter pays" principle the costs of cleaning up abandoned hazardous-waste sites are borne by the "responsible parties" and not by the taxpayers; from the standpoint of cost-benefit analysis, with its emphasis on efficiency, who pays for a cleanup is a second-order issue.

The anthropologist Mary Douglas is the leading proponent of the view that cultural anthropology provides a useful lens through which to view the social influences and cultural values that underlie risk perception. Douglas (1985 p. 3) argues that risk perception is a social phenomenon:

The professional discussion of cognition and choice has no sustained theorizing about the social influences which select particular risks for public attention. Yet it is hard to maintain seriously that perception of risk is private With no link between cultural analysis and cognitive science, clashes inevitably occur between theory and evidence. Since the theory is not being radically adjusted, irrationality tends to be invoked to protect the too narrow definition of rationality. So instead of a sociological, cultural, and ethical theory of human judgment, there is an unintended emphasis on perceptual pathology.

Margolis (1996) is firmly in the "perceptual pathology" camp, arguing that people often do not understand why they believe what they believe and that, under some circumstances, "habits of mind" (e.g., framing and loss aversion) cause the public to misperceive reality. Like Breyer, Margolis believes that the experts provide reliable assessments of the risks, but that the public mistakenly rejects these expert assessments. Margolis attributes these rejections to cognitive illusions analogous to visual or perceptual illusions such as the famous Muller-Lyer illusion (Margolis, 1996 p. 58) and Koehler's duck/rabbit (p. 74).³

II. What To Do When Perceptions Differ

Margolis (1996 p. 161) touches briefly and inconclusively on the issue of how to deal with risk in a democratic society when the public's fears are not shared by the experts: "If enough people feel worried about some risk, however

³ Margolis's subtitle, "Why the Public and the Experts Disagree on Environmental Issues," conveys the book's focus more accurately than its title, *Dealing with Risk*.

remote and cautiously calculated, then it makes sense to say that the government ought to respond to that. How to respond is less clear." Margolis, however, would clearly oppose spending money on an expensive water treatment system just because "people feel worried about some risk," when the experts "know" that the public's worries are unfounded. His dismissal of "psychic benefits" ("People who were worried feel protected, and so feel better even in a case where it was really true that there was nothing to be worried about in the first place" [Margolis, 1996 pp. 197–98]) is similar to Breyer's.

As Margolis acknowledges, "How to respond is less clear." But if we follow Margolis and interpret the problem as a psychic one—and, more specifically, one reflecting "perceptual pathology"—then the set of possible responses is small. "Education" and "risk communication," in Happyville and elsewhere, have been less than completely successful, and pharmacological responses (e.g., dispensing Prozac) raise uncomfortable ethical issues.

Both Breyer and Margolis believe that the priorities of regulatory agencies are unduly influenced by public misperceptions of risks. Breyer sees reforming and depoliticizing the regulatory process—a solution that harnesses what he calls "the inherent virtues of civil service to bring about improved performance" (Breyer, 1993 p. 67), drawing on "the virtues of bureaucracy" (pp. 59–61)—as the best chance for improving government risk regulation. Following Breyer, Margolis (1996 p. 177) advocates creating a procedure that "insulates an administrator (or legislator) from attack for an assessment that fails to come to the immediately popular conclusion." Thus, Margolis concludes: "What seems to be needed is a voice insulated from the political and public relations difficulties that inhibit government officials from fulfilling what might be seen as directly their responsibility" (p. 203).

These proposals for bureaucratic reform will trouble critics who, in the tradition of John Dewey, take democratic political ideals seriously, as well as those who are concerned that regulators are vulnerable to "capture." The proposals seem designed to

protect regulators from capture by Congress and the public, yet regulators who are insulated from scrutiny by Congress and the public may be more vulnerable to capture by the industries and interests they are supposed to regulate.

III. Sorting Out the Questions

The divergence between the perceptions of the public and those of experts poses at least five distinct but related questions:

- (i) The first belongs to the philosophy of science and is epistemological. How can scientists infer risk-related facts (e.g., how much would reducing the occupational exposure standard for benzene from 10 ppm to 1 ppm reduce the incidence of leukemia 25 years in the future?) given the difficulties of drawing such inferences from bioassay and epidemiological data?⁴
- (ii) The second belongs to science studies or the sociology of science. Who is an expert? How are expert perceptions actually formed?⁵

⁴ Pollak (1995) discusses the difficulties of dose/response assessment using formaldehyde as an example and provides references to the literature. Breyer (1993 pp. 42–43) candidly acknowledges the difficulties of risk assessment:

Predicting risk is a scientifically related enterprise, but it does not involve scientists doing what they do best, namely developing theories about how x responds to y , other things being equal Moreover, where prediction involves a weak relationship, such as that between a small dose of a substance and a later cancer death, as well as long lead times, such as exposure for twenty years or more, it is difficult or impossible for predictors to obtain empirical feedback, which is necessary (for them as for all of us) to confirm or correct their theories.

John Wargo (1996) provides an extended discussion of exposure assessment in the context of pesticide regulation, emphasizing the importance of focusing on those who may have unusually high exposure (e.g., although apple sauce and apple juice are minor components of adults' diets, they are major components of infants' diets).

⁵ See, for example, Andrew Pickering (1995) and the papers collected in Pickering (1992).

- (iii) The third belongs to social psychology and sociology. How are public (i.e., non-expert) perceptions formed? What roles do the media play? What roles do "public-interest," "private-interest," and "special-interest" groups play? What roles do government pronouncements and actions play?
- (iv) The fourth belongs to normative political philosophy. How should governments regulate risks when the perceptions of the public diverge from those of the experts? What weight, if any, should be accorded public (mis)perceptions? Should public concerns receive greater weight when expert opinion is divided? Should public concerns receive greater weight when the experts recognize that the uncertainty associated with a particular risk assessment is great?
- (v) The fifth belongs to positive political science. How do governments regulate risks when the perceptions of the public diverge from those of the experts? What role does the median voter play in this process? What role do experts play? What roles do risk analysis and cost-benefit analysis play?⁶

This final question deserves much more attention than it has thus far received. The literatures on risk analysis and cost-benefit analysis are primarily normative, focusing on what governments "ought" to do and whether they have done the "right" thing. Thus, they neglect positive questions such as: how do governments actually behave? Which risks do governments regulate and which do they not regulate? Which instruments and procedures do governments use to regulate risks? To what extent do differences in the choices made (e.g., between the Clean Air Act and the Safe Drinking Water Act) reflect technical differences in the risks and the technologies available to control airborne and waterborne hazards, and

to what extent do they reflect political, legal, social, and cultural factors?

Even if the notion of risk-related facts were not problematic, it would not be obvious how democratic governments ought to act when the public and the experts perceive risks differently. Public fears and (mis)perception, after all, are themselves stubborn risk-related facts—facts that the public does not easily surrender, even after "education" by "risk communication." (Recall that, as Director of Environmental Protection in Happyville, your skillful efforts to communicate the experts' confidence in the water supply failed to reassure the public.) Public fears clearly play a role in determining government policies regulating risks, and perhaps they should. Utilitarians—and most welfare economists and policy analysts approach public policy from a utilitarian perspective—should consider whose beliefs (the public's or the experts') should be used to calculate expected benefits.

REFERENCES

- Breyer, Stephen. *Breaking the vicious circle: Toward effective risk regulation*. Cambridge, MA: Harvard University Press, 1993.
- Douglas, Mary. *Risk acceptability according to the social sciences*. New York: Russell Sage Foundation, 1985.
- Margolis, Howard. *Dealing with risk: Why the public and the experts disagree on environmental issues*. Chicago: University of Chicago Press, 1996.
- Pickering, Andrew, ed. *Science as practice and culture*. Chicago: University of Chicago Press, 1992.
- . *The mangle of practice: Time, agency, and science*. Chicago: University of Chicago Press, 1995.
- Pollak, Robert A. "Regulating Risks." *Journal of Economic Literature*, March 1995, 33(1), pp. 179–91.
- . "Government Risk Regulation." *Annals of the American Academy of Political and Social Science*, May 1996, 545, pp. 25–34.

⁶Theodore Porter (1995) provides an interesting discussion of the history and politics of cost-benefit analysis.

- _____. "Risk Regulation," in Peter Newman, ed., *The new Palgrave dictionary of economics and the law*. London: Macmillan, 1998 (forthcoming).
- Porter, Theodore M. *Trust in numbers: The pursuit of objectivity in science and public life*. Princeton, NJ: Princeton University Press, 1995.
- Portney, Paul R. "Trouble in Happyville." *Journal of Policy Analysis and Management*, Winter 1992, 11(1), pp. 131-32.
- Wargo, John. *Our children's toxic legacy: How science and law fail to protect us from pesticides*. New Haven: Yale University Press, 1996.

General-Equilibrium Treatment Effects: A Study of Tuition Policy

By JAMES J. HECKMAN, LANCE LOCHNER, AND CHRISTOPHER TABER*

This paper considers the effects of changes in tuition on schooling and earnings; accounting for general-equilibrium effects on skill prices. The typical evaluation estimates the response of college enrollment to tuition variation using geographically dispersed cross sections of individuals facing different tuition rates. These estimates are then used to determine how subsidies to tuition will raise enrollment. The impact of tuition policies on earnings is evaluated using a schooling-earnings relationship fit on pre-intervention data and does not account for the enrollment effects of the taxes raised to finance the tuition subsidy. Thomas Kane (1994) exemplifies this approach.

The danger in this widely used practice is that what is true for policies affecting a small number of individuals need not be true for policies that affect the economy at large. A national tuition-reduction policy that stimulates substantial college enrollment will likely reduce college skill prices, as advocates of the policy claim. However, agents who account for these changes will not enroll in school at the levels calculated from conventional procedures, which ignore the impact of the induced enrollment on earnings. As a result, standard policy-evaluation practices are likely to be misleading about the effects of tuition policy on schooling attainment and wage inequality. The empirical question is: how misleading? We show that these practices lead to estimates of enrollment responses that are more than ten times larger than the long-run general-equilibrium

effects. We also improve on current practice in the treatment-effects literature by considering both the gross benefits of the program and the tax costs of financing the treatment as borne by different groups.

Evaluating the general-equilibrium effects of a national tuition policy requires more information than the tuition-enrollment parameter that is the centerpiece of partial-equilibrium policy analysis. Most policy proposals extrapolate well outside the range of known experience and ignore the effects of induced changes in skill quantities on skill prices. To improve on current practice, we have developed an empirically justified, dynamic, overlapping-generations, general-equilibrium framework for the pricing of heterogeneous skills. It is based on an empirically grounded theory of the supply of schooling and post-school human capital, where different schooling levels represent different skills. Individuals differ in learning ability and in initial endowments of human capital. Household saving behavior generates the aggregate capital stock, and output is produced by combining the stocks of different human capitals with physical capital. The framework explains the pattern of rising wage inequality experienced in the United States in the past 30 years (Heckman et al., 1998). In this paper we apply this framework to evaluate tuition policies that attempt to increase college enrollment.

The statistical and econometric literature on "treatment effects" is remarkable for its inattention to the market consequences of the programs it evaluates. The widely used "Rubin" model (Donald Rubin, 1978) assumes no interactions among the agents being analyzed. The paradigm in the econometric literature on treatment effects is that of evaluating the effectiveness of a drug. It assumes that there are no spillovers to society at large that flow from drug use (or "treatment") by individuals.

* Heckman and Lochner: Department of Economics, University of Chicago, 1126 E. 59th Street, Chicago, IL 60637; Taber: Department of Economics, Northwestern University, 2003 Sheridan Road, Evanston, IL 60208. We thank Lars Hansen and Alice Nakamura for helpful comments. This research was supported by a grant from the Russell Sage Foundation and by NSF grant SBR-93-21-048.

The literature in economics recognizes these spillover effects. The classical analysis of union relative-wage effects by H. Gregg Lewis (1963) explicitly accounts for the discrepancy between the effects of treatment (unionism) on an individual and treatment applied to an industry when prices adjust to industry-wide unionization levels. Our analysis extends Lewis's static general-equilibrium framework to a dynamic setting with skill formation.

I. Conventional Models of Treatment Effects

The standard framework for a micro-econometric program evaluation is partial-equilibrium in character (see Heckman and Richard Robb, 1985). For a given individual i , $Y_{0,i}$ is defined to be the outcome the individual experiences if he does not participate in the program, and $Y_{1,i}$ is the outcome he experiences if he does participate. The treatment effect for person i is $\Delta_i = Y_{1,i} - Y_{0,i}$. When interventions have general-equilibrium consequences, these effects depend on who else is treated and the market interaction between the treated and the untreated.

To see the problems that arise in the standard framework, consider instituting a national tuition policy. In this case, $Y_{0,i}$ is person i 's wage if he does not attend college, and $Y_{1,i}$ is his wage if he does attend. The "parameter" Δ_i then represents the impact of college, and it can be used to estimate the impact of tuition policies on wages. It is a constant, or policy-invariant, parameter only if wages ($Y_{0,i}$, $Y_{1,i}$) are invariant to the number of college and high-school graduates in the economy.

In a general-equilibrium setting, an increase in tuition increases the number of individuals who attend college, which in turn decreases the relative wages of college attendees, $Y_{1,i}/Y_{0,i}$. In this case, the program not only impacts the wages of individuals who are induced to move by the program, but also has an impact on the wages of those who do not. For two reasons, then, the "treatment-effect" framework is inadequate. First, the parameters of interest depend on who in the economy is "treated" and who is not. Second, these parameters do not measure the full impact of the program. For example, increasing tuition subsidies may increase the earnings of uneducated

individuals who do not take advantage of the subsidy. To pay for the subsidy, the highly educated would be taxed, and this may affect their investment behavior. In addition, more competitors for educated workers enter the market as a result of the policy, and their earnings are depressed. Conventional methods ignore the effect of the policy on nonparticipants. In order to account for these effects, it is necessary to conduct a general-equilibrium analysis.

II. Exploring Increases in Tuition Subsidies in a General-Equilibrium Model

We first simulate the effects of a revenue-neutral \$500 increase in tuition subsidy (financed by a proportional tax) on enrollment in college and wage inequality starting from a baseline economy that describes the United States in the mid-1980's and that produces wage growth profiles and schooling-enrollment and capital-stock data that match micro and macro statistics. The partial-equilibrium increase in college attendance is 5.3 percent. This analysis holds skill prices and, therefore, college and high-school wage rates fixed—a typical assumption in micro-economic treatment-effect analyses.

When the policy is evaluated in a general-equilibrium setting, the estimated effect falls to 0.46 percent. Because the college-high-school wage ratio falls as more individuals attend college, the returns to college are less than when the wage ratio is held fixed. Rational agents understand this effect of the tuition policy on skill prices and adjust their college-going behavior accordingly. Policy analysis of the type offered in the treatment-effect literature ignores the responses of rational agents to the policies being evaluated. There is substantial attenuation of the effects of tuition policy on capital and the stocks of the different skills in our model. In our baseline specification, we allow skill prices and interest rates to adjust in general equilibrium but hold the pre-subsidy tuition level fixed. Simulating the policy under a number of additional alternative assumptions about the parameters of the economic model, including a case where tuition costs rise with enrollment, reproduces the basic result of substantial partial-

equilibrium effects and much weaker general-equilibrium effects.

Our steady-state results are long-run effects. When we simulate the model with rational expectations, the short-run enrollment effects are also very small, as agents anticipate the effects of the policy on skill prices and calculate that there is little gain from attending college at higher rates. If we simulate using myopic expectations, the short-run enrollment effects are much closer to the estimated partial-equilibrium effects. All of these results are qualitatively robust to the choice of different tax schedules. Progressive tax schedules choke off skill investment and lead to lower enrollment responses in general equilibrium.

We next consider the impact of a policy change on discounted earnings and utility. We decompose the total effects into benefits and costs, including tax costs for each group. For the sake of brevity, we report overall results, and not the results by ability type. Table 1 compares outcomes in two steady states: (i) the benchmark steady state and (ii) the steady state associated with the new tuition policy. Given that the estimated general-equilibrium schooling response to a \$500 subsidy is small, we instead use an extremely high \$5,000 subsidy for the purpose of exploring general-equilibrium effects. The row "high-school-high-school" reports the change in a variety of outcome measures for those persons who would be in high school under the benchmark or new policy regime; the high-school-college row reports the change in the same measures for high-school students in the benchmark who are induced to attend college only by the new policy; college-high-school outcomes refer to those persons in college in the benchmark economy who only attend high school after the new policy is put in place; and so forth.

By the measure of the present value of earnings, some of those induced to change are worse off. Contrary to the monotonicity assumption built into the LATE parameter of Guido Imbens and Joshua Angrist (1994), defined in this context as the effect of tuition change on the earnings of those induced to go to college, we find that the tuition policy produces a two-way flow. Some people who would have attended college in the benchmark regime no longer do so. Again, contrary to the

TABLE 1—SIMULATED EFFECTS OF A \$5,000 TUITION SUBSIDY ON DIFFERENT GROUPS: STEADY-STATE CHANGES IN PRESENT VALUE OF LIFETIME WEALTH (THOUSANDS OF 1995 DOLLARS)

Group (proportion) ^a	(i) After-tax earnings using base tax	(ii) After-tax earnings	(iii) After-tax earnings net of tuition	(iv) Utility
High-school-high-school (0.528)	9.512	-0.024	-0.024	-0.024
High-school-college (0.025)	-4.231	-13.446	1.529	1.411
College-high-school (0.003)	-46.711	-57.139	-53.019	-0.879
College-college (0.444)	-7.654	-18.204	0.420	0.420

Notes: Column (i) reports the after-tax present value of earnings in thousands of dollars discounted using the after-tax interest rate, where the tax rate used for the second steady state is the base tax rate. Column (i) reports just the effect on earnings; column (ii) adds the effect of taxes; column (iii) adds the effect of tuition subsidies; and column (iv) includes the nonpecuniary costs of college, expressed in dollars.

^a The groups denote counterfactual groups. For example, the high-school-high-school group consists of individuals who would not attend college in either steady state, and the high-school-college group would not attend college in the first steady state, but would in the second, etc.

implicit assumption built into LATE that only those who change status are affected by the policy, the rest of society also is affected by the policy. People who would have gone to college without the policy and continue to do so after the policy is put in place are financially worse off for two reasons: (i) the price of their skill is depressed, and (ii) they must pay higher taxes to finance the policy. However, they now receive a tuition subsidy, and for this reason, on net, they are slightly better off both financially and in terms of utility. Those who would abstain from attending college in both steady states are essentially indifferent between them. They pay higher taxes, but their skill becomes more scarce, and their wages rise. Those induced to attend college by the

policy are better off in terms of utility but are not better off in terms of income. Note that neither category of non-changers is a natural benchmark for a "difference in differences" estimator. The movement in their wages before and after the policy is due to the policy and cannot be attributed to a benchmark "trend" that is independent of the policy.

Table 2 presents the impact of the \$5,000 tuition policy on the log earnings of individuals with ten years of work experience for different definitions of treatment effects. The partial-equilibrium version given in the first column holds skill prices constant at initial steady-state values. The general-equilibrium version given in the second column allows prices to adjust when college enrollment varies. Consider four parameters initially defined in a partial-equilibrium context. The *average treatment effect* is defined for a randomly selected person in the population in the benchmark economy and indicates how that person would gain in wages by moving from high school to college. The parameter *treatment on the treated* is defined as the average gain over the noncollege alternative of those who attend college. The parameter *treatment on the untreated* is defined as the average gain over the college wage received by individuals who did not attend college. The *marginal treatment effect* is defined for individuals who are indifferent between going to college or not. It is a limit version of the LATE parameter under conventional assumptions made in discrete-choice theory (Heckman, 1997). Taber (1997) considers this parameter in his analysis of schooling choices. Column (ii) presents the general-equilibrium version of these treatment effects. *Treatment on the treated* compares the earnings of college graduates in the benchmark economy with what they would earn if no one went to college.¹ *Treatment on the un-*

TABLE 2—TREATMENT-EFFECT PARAMETERS, PARTIAL-EQUILIBRIUM AND GENERAL-EQUILIBRIUM (GE) DIFFERENCE IN LOG EARNINGS: COLLEGE GRADUATES (COL) VERSUS HIGH-SCHOOL GRADUATES (HS) WITH TEN YEARS OF WORK EXPERIENCE

Parameter	(i) Prices fixed	(ii) Prices vary	(iii) Percentage of sample
Average treatment effect (ATE)	0.281	1.801	100
Treatment on treated (TT)	0.294	3.364	44.7
Treatment on untreated (TOU)	0.270	-1.225	55.3
Marginal treatment effect (MTE)	0.259	0.259	—
LATE, \$5,000 subsidy:			
Partial equilibrium	0.255	—	23.6
GE (HS→COL)	0.253	0.227	2.48
LATE			
GE (COL→HS)	0.393	0.365	0.34
LATER			
GE net TLATE	—	0.244	2.82
LATE \$500 subsidy:			
Partial equilibrium	0.254	—	2.37
GE (HS→COL)	0.250	0.247	0.24
LATE			
GE (COL→HS)	0.393	0.390	0.03
LATER			
GE net TLATE	—	0.264	0.27

Notes: (i) "Prices fixed" denotes the difference in log earnings between college and high-school graduates for various groups. Prices are held constant at their initial steady-state levels when wage differences are calculated. In column (ii) we allow prices to adjust in response to the change in schooling proportions when calculating wage differences. For each row, column (iii) presents the total fraction of the sample over which the parameter is defined. The LATE group denotes the effect on earnings for persons who would be induced to attend college by a tuition change. In the case of GE, LATE measures the effect on individuals induced to attend college when skill prices adjust in response to quantity movements among skill groups. The partial-equilibrium LATE measures the effect of the policy on those induced to attend college when skill prices are held constant at the benchmark level.

¹ In the empirical general-equilibrium model of Heckman et al. (1998), Inada conditions for college and high school are not imposed, and the marginal product of each skill group when none of it is utilized is a bounded number. If Inada conditions were imposed, and the marginal product of a skill goes to infinity when the aggregate quantity of skill in the economy goes to zero, the counterfactual and the counterfactual treatment on the untreated would not be defined.

treated is defined analogously by comparing what high-school graduates in the benchmark economy would earn if everyone in the population were forced to go to college. The *average treatment effect* compares the average

earnings in a world in which everyone attends college against the earnings in a world in which nobody attends college. Such dramatic policy shifts produce large estimated effects. In contrast, the general-equilibrium marginal treatment-effect parameter considers the gain to attending college for people on the margin of indifference between attending college and attending high school. In this case, as long as the mass of people in the indifference set is negligible, the partial- and general-equilibrium parameters are the same.

The final set of parameters we consider are versions of the LATE parameter. This parameter depends on the particular intervention being studied and its magnitude. The partial-equilibrium version of LATE is defined for the outcomes of individuals induced to attend college, assuming that skill prices do not change. The general-equilibrium version is defined for the individuals induced to attend college when prices adjust in response to the policy. The two LATE parameters are quite close to each other and are also close to the marginal treatment effect.² General-equilibrium effects change the group over which the parameter is defined compared to the partial-equilibrium case. For the \$5,000 subsidy, there are substantial price effects, and the partial-equilibrium parameter differs substantially from the general-equilibrium parameter.

We also present partial- and general-equilibrium estimates for two extensions of the LATE concept: LATER (the effect of the policy on those induced to drop out of college and go to high school; reverse LATE) and TLATE (the effect of the policy on all of those induced to change whichever direction they flow). LATER is larger than LATE, indicating that those induced to drop out of college have larger gains from dropping out than those induced to enter college have from entering. TLATE is a weighted average of LATE and LATER, with weights given by the relative

proportion of people who switch in each direction.

III. Summary

This paper defines and estimates general-equilibrium treatment effects. Focusing on the impact of tuition policy, we find that general-equilibrium impacts of tuition on college enrollment are an order of magnitude smaller than those reported in the literature on microeconomic treatment effects. The assumptions used to justify the LATE parameter in a microeconomic setting do not carry over to a general-equilibrium framework. Policy changes, in general, induce two-way flows and violate the monotonicity (or one-way flow) assumption of LATE. We extend the LATE concept to allow for the two-way flows induced by the policies. We present a more comprehensive approach to program evaluation by considering both the tax and benefit consequences of the program being evaluated and placing the analysis in a market setting.

REFERENCES

- Heckman, James. "Instrumental Variables: A Study of Implicit Behavioral Assumptions in One Widely Used Estimator." *Journal of Human Resources*, Summer 1997, 32(3), pp. 441-62.
- Heckman, James; Lochner, Lance and Taber, Christopher. "Explaining Rising Wage Inequality: Explorations with a Dynamic General Equilibrium Model of Labor Earnings with Heterogeneous Agents." *Review of Economic Dynamics*, 1998 (forthcoming).
- Heckman, James J. and Robb, Richard, Jr. "Alternative Methods for Estimating the Impact of Interventions," in James J. Heckman and Burton Singer, eds., *Longitudinal analysis of labor market data*. Cambridge: Cambridge University Press, 1985, pp. 156-245.
- Imbens, Guido and Angrist, Joshua. "Identification and Estimation of Local Average

² The latter is a consequence of the discrete-choice framework we use to model schooling choices in our model (see Heckman, 1997).

- Treatment Effects." *Econometrica*, March 1994, 62(2), pp. 467-75.
- Kane, Thomas. "College Entry by Blacks Since 1970: The Role of College Costs, Family Background, and the Returns to Education." *Journal of Political Economy*, October 1994, 102(5), pp. 878-911.
- Lewis, H. Gregg. *Unionism and relative wages*. Chicago: University of Chicago Press, 1963.
- Rubin, Donald. "Bayesian Inference for Causal Effects: The Role of Randomization." *Annals of Statistics*, January 1978, 6(1), pp. 34-58.
- Taber, Christopher. "The Rising College Premium in the Eighties: Returns to College or Returns to Ability?" Unpublished manuscript, Northwestern University, 1997.

GOVERNMENT IN TRANSITION[†]

Regulatory Discretion and the Unofficial Economy

By SIMON JOHNSON, DANIEL KAUFMANN, AND PABLO ZOIDO-LOBATÓN*

Politicization of economic activity means the exercise of control rights over firms by politicians and bureaucrats. In most countries politicians maintain property rights in firms, typically in the form of residual control rights as defined by Sanford Grossman and Oliver Hart (1986). These control rights may have served an ideological agenda in the past, but they are often used to further the private agenda of politicians and bureaucrats. A recent literature has established the presence of these problems in countries as diverse as Peru, France, Russia, and Ukraine (Hernando de Soto, 1989; Andrei Shleifer and Robert Vishny, 1993, 1994; Kaufmann and Paul Siegelbaum, 1997; Shleifer, 1997). But how widespread are these rights and how damaging are their effects around the world?

The usual presumption in the economics literature is that a predatory government simply leads to lower total economic activity, but for Eastern Europe and the former Soviet Union since 1989, Johnson et al. (1997) showed that businesses have responded to politicization by going "underground." Instead of registering their activities, managers prefer not to pay taxes and not to benefit from key publicly pro-

vided services, such as legal enforcement of contracts. For these economies in transition from communism, there is evidence of a downward spiral, in which firms leaving the official sector reduce state revenue, which reduces publicly provided services and further reduces the incentive to register in the official sector.¹ Most of the former Soviet Union has thus ended up in a "bad" equilibrium with low tax revenue, high unofficial economy as a percentage of GDP, and low quality of publicly provided services.

This previous work on transition economies suggests that, while formal rules may count in some instances, what really matters is how regulations and tax rules are actually implemented. If the rules are fine on paper but officials have a great deal of discretion in their interpretation and implementation, this leads to a higher effective burden on business, more corruption, and a greater incentive to move to the unofficial economy. This general idea leads to three specific propositions. First, the share of the unofficial economy in GDP should be higher when there is more regulation and more discretion for officials regarding how the regulatory system operates. Second, the unofficial economy should be larger when there is a bigger tax burden on firms in the official sector, where "burden" on the firm is the outcome of how the tax system is administered as well as what the rates are. Third, a larger unofficial economy should be correlated with weaker publicly provided services, as measured by corruption and the "rule of law" (particularly the legal protection provided to private-sector business investments).

¹ Norman Loayza (1996) has similar theoretical results for Latin America. In his model, unregistered firms use but do not pay for public services, thus leading to congestion costs for public goods, such as roads, and lower growth.

[†] *Discussants:* Jean-Laurent Rosenthal, University of California—Los Angeles; Yingyi Qian, Stanford University; Avner Greif, Stanford University.

* Johnson: Sloan School of Management, Massachusetts Institute of Technology, 50 Memorial Drive, Cambridge, MA 02142-1347; Kaufmann and Zoido-Lobaton: World Bank, 1818 H Street, N.W., Washington, DC 20433. Johnson gratefully acknowledges support from the Entrepreneurship Center at MIT. We thank Kenneth Sokoloff, Jean-Laurent Rosenthal, Andrei Shleifer, and Normal Loayza for discussions and suggestions. The authors are responsible for the paper's views, errors, and omissions. Views expressed do not necessarily reflect those of the affiliated institutions. The presentation of indexes here does not constitute an endorsement by the authors or their affiliated institutions of any individual country rating.

This paper finds support for these propositions in a broad set of countries for which there exist at least roughly comparable estimates of the unofficial economy in the 1990's. We have measures for the unofficial economy for 49 countries in three regions of the world: Latin America, the OECD, and the former Soviet bloc. A different methodology is used for each region, but the numbers appear to be comparable; see Johnson et al. (1998) for the detailed estimates. The sample for our regressions varies between 32 and 49 countries, depending on the coverage of right-hand-side variables. We have not found comparable data for the unofficial economy in East Asia or for Africa, so these countries are excluded from the regressions. We use Brazil and Russia as illustrative regional benchmarks throughout and also report on OECD-specific countries where relevant.

I. Regulation and Bureaucracy

The Heritage Foundation's measure of regulation is higher, on a scale of 1 to 5, for countries that had regulations that are worse for business in 1996 (Bryan Johnson and Thomas Sheehy, 1997). This measure includes both the formal rules and the way they are enforced. The Czech Republic actually receives the top score; it is the only country in our sample to get a perfect 1. Most OECD countries score 2. Russia scores 4, while Brazil scores 3. Table 1 shows that a one-point increase in this index is associated with a 14.7-percentage-point increase in the share of the unofficial economy. Controlling for log GDP per capita reduces the coefficient on the regulation variable to 8.1, but it remains significant.

The Global Competitiveness Survey reports results from a 1997 survey of executives on the extent of regulatory discretion and lax enforcement of rules, on a scale of 1 to 7 (World Economic Forum, 1997). Russia has the lowest score of 2.01, while Brazil rates better with 3.46. Most of the OECD countries score 4.5 or higher; Switzerland has the highest score with 5.64 in our sample. Singapore had the highest score worldwide in the survey, with 6.36. Table 1 shows that a one-point-higher score for this index is correlated with a 9.2-percentage-point fall in the share of the unof-

TABLE 1—REGRESSIONS OF UNOFFICIAL ECONOMY (AS PERCENTAGE OF GDP) ON MEASURES OF REGULATION

Independent variable	Regression					
	(i)	(ii)	(iii)	(iv)	(v)	(vi)
Log GDP per capita		-7.4* (1.6)		-7.4* (2.3)		-1.0 (2.9)
Measures of regulation						
Regulation ^a	14.7* (2.5)	8.1* (2.6)				
Regulatory discretion ^a			-9.2* (1.7)	-2.9 (2.5)		
Bureaucratic quality ^a					-8.5* (1.0)	-7.7* (2.3)
R ² :	0.43	0.62	0.47	0.60	0.65	0.65
Number of observations:	47	47	34	34	39	39
Independent variable	Regression					
	(vii)	(viii)	(ix)	(x)	(xi)	(xii)
Log GDP per capita		-7.4* (2.0)		-7.3* (1.5)		-7.0* (1.6)
Measures of regulation (continued)						
Economic freedom ^a	-2.5* (0.5)	-0.8 (0.6)				
Measure of taxation						
Tax burden ^a			-11.7* (2.4)	-6.5* (2.1)		
Tax rules ^a					3.5* (0.7)	1.9* (0.7)
R ² :	0.38	0.54	0.43	0.68	0.37	0.57
Number of observations:	42	42	34	34	42	42

Notes: Standard errors are in parentheses.

^a A higher value for this variable stands for a better score for private business.

^b A higher value for this variable stands for a worse score for private business.

* Statistically significant at the 5-percent level.

ficial economy. However, this measure is not significant once we control for log GDP per capita.

The 1997 *International Country Risk Guide* (Political Risk Services, 1997) measures expert opinion of "bureaucratic quality" on a scale of 1 to 6, where a higher score means that bureaucrats operated in a more efficient and predictable way between 1990 and 1997. Guatemala and Panama have the lowest score of 1.44; Russia scores 3.19; and Brazil scores 4.0. The best OECD countries, such as the United Kingdom score 6.0. Table 1 shows that

a one-point increase in this index implies an 8.5-percentage-points decrease in the share of the unofficial economy. Controlling for log GDP per capita reduces the coefficient only slightly to -7.7 , and it remains highly significant.

Freedom House's 1995–1996 measure of economic freedom is higher for countries with "better" regulation (i.e., more pro-business), on a scale of 0 to 16 (Richard E. Messick, 1996). The United Kingdom, the United States, Denmark, Sweden, and Holland tie for top position with a score of 16, while Azerbaijan has the lowest score of 1. Russia and Brazil score 7. Table 1 shows that a one-point increase in this scale is associated with a 2.5-percent fall in the share of the unofficial economy, but this coefficient loses significance when we control for GDP per capita.

In summary, we find strong evidence that less regulation (i.e., a regulatory regime that is more business-friendly and presumably represents less political control rights) is correlated with a lower share of the unofficial economy. However, countries with a higher income level also have a lower level of the unofficial economy, so when we control for income level two out of four regulation variables become insignificant at the 5-percent level. The effect of bureaucratic quality and the way regulations are administered appear to be particularly strong. This supports the idea that regulatory discretion is an important cause of unofficial activity.

II. Taxation

The 1997 Global Competitiveness Survey rates tax burden from the firm's standpoint; a higher score was given when executives considered the tax system to be better for business, on a scale of 1 to 7 (World Economic Forum, 1997). This measure captures not just tax rates, but also the way the tax system is administered (e.g., if tax officials abuse higher levels of discretion, this would likely translate into a worse score). Ukraine has the lowest score in our sample, with 1.58, and the United Kingdom has the highest score, with 4.60. Russia scores 1.80, and Brazil scores 2.22. A one-point increase in this variable reduces the share of the unofficial economy by 11.7 per-

centage points. Controlling for log GDP per capita reduces the coefficient to -6.5 but it remains significant.

The Fraser Institute measure of top marginal tax rates is higher for countries that had lower tax rates, on a scale of 1–10, in 1995 (James Gwarney and Robert Lawson, 1997). In this case the index captures formal rates, but not the way the system is administered. The "best" tax rates are in seemingly unlikely places: Bolivia and Uruguay both score a perfect 10.² The worst (i.e., highest) tax rates are in Italy, Belgium, Sweden, Denmark, and Romania, all of which score the lowest attainable value of 1. The United States scores 7, and the United Kingdom scores 5, while Russia and Brazil both score 8. Chile scores 4, which is the best in Latin America. Table 1 shows that a one-point increase in this index is actually associated with a 3.5-percentage-point increase in the share of the unofficial economy (i.e., countries with lower marginal tax rates actually have a larger share of the unofficial economy). Controlling for log GDP per capita reduces the coefficient on this index to 1.9, but it remains significant.

The contrast between the results of these two tax variables points to the importance of how the tax and regulatory system operates, rather than the nature of the formal rules. Countries with high marginal tax rates but a low tax burden (as evaluated by executives) actually have a low share of the unofficial economy as a percentage of GDP (e.g., Scandinavia; see Fig. 1). Russia has relatively low marginal tax rates but was rated with a high tax burden because of the way the tax system operates, and thus it is associated with a relatively high share of the unofficial economy in GDP.

III. Rule of Law and Corruption

Political Risk Services' 1997 *International Country Risk Guide* contains a "rule-of-law index" which is higher where the law-and-order tradition was stronger during 1990–1997, on a scale of 0–6. The United States

² Bolivia's recent tax reform is presumably reflected in this rating.

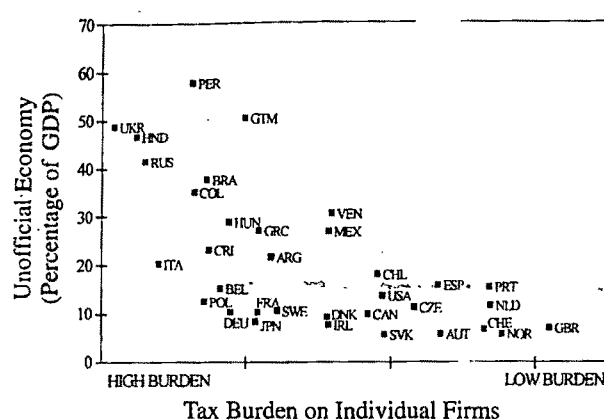


FIGURE 1. UNOFFICIAL ECONOMY AND TAX BURDEN ON INDIVIDUAL FIRMS

Notes: Unofficial-economy estimates are from Johnson et al. (1998); the tax burdens on individual firms are from World Economic Forum (1997).

and several other OECD countries achieve the highest level of 6. In our sample, Colombia has the lowest score of 1.4. Russia scores 3.5, and Brazil scores 3.4. Table 2 shows that a one-point increase in the value of this index is associated with a 10.6-percentage-point fall in the share of the unofficial economy. In this case log GDP per capita is not significant, and including this control variable reduces the estimated coefficient on the index only to -9.3.

The Heritage Foundation's index of property rights is lower where property rights were more secure, on a scale of 1-5, in 1996 (Johnson and Sheehy, 1997). The only non-OECD country to score a perfect 1 is Chile. Four previously communist countries have the worst score of 4: Romania, Ukraine, Georgia, and Azerbaijan. Russia and Brazil score 3. Table 2 shows that a one-point increase in this index is associated with a 13.4-percent increase in the share of the unofficial economy. Controlling for log GDP per capita reduces the coefficient to 8.0, but it remains significant.

In the Fraser Institute measure of "Equality of Citizens Under the Law and Access of Citizens to a Non-discriminatory Judiciary," a higher score means a "better" legal system in 1995, on a scale of 0-10 (Gwarney and Lawson, 1997). Only Belgium, Holland, Sweden, Norway, Denmark, and Switzerland get the top score of 10. Italy, the United Kingdom, and the United States score 7.5. Russia scores

TABLE 2—REGRESSION OF UNOFFICIAL ECONOMY (AS PERCENTAGE OF GDP) ON LEGAL ENVIRONMENT AND CORRUPTION

Independent variable	Regression					
	(i)	(ii)	(iii)	(iv)	(v)	(vi)
Log GDP per capita		-1.9 (1.7)		-4.8 [†] (2.6)		-5.2* (1.9)
Legal environment						
ICRG rule-of-law index, 1990-1997 ^a		-10.6* (1.0)	-9.3* (1.5)			
Property rights ^b				13.4* (1.8)	8.0* (3.4)	
Equality of citizens before the law ^a						-3.8* (0.6)
						-2.3* (0.8)
R ² :	0.77	0.78	0.55	0.58	0.53	0.60
Number of observations:	39	39	47	47	43	43
Independent variable	Regression					
	(vii)	(viii)	(ix)	(x)	(xi)	(xii)
Log GDP per capita		-4.0* (2.3)		-5.8* (2.5)		-6.5* (1.9)
Corruption						
Transparency International (extended) ^a		-5.1* (0.7)	-3.5* (1.1)			
World Economic Forum ^a				-8.0* (1.3)	-3.9* (2.1)	
Impulse's exporter bribery index ^b						1.7* (0.4)
						0.8 [†] (0.4)
R ² :	0.57	0.60	0.55	0.62	0.36	0.50
Number of observations:	43	43	34	34	44	44

Notes: Standard errors are in parentheses.

^a A higher value for this variable stands for a better score for private business.

^b A higher value for this variable stands for a worse score for private business.

[†] Statistically significant at the 10-percent level.

* Statistically significant at the 5-percent level.

2.5, and Brazil scores 0.³ Table 2 shows that a one-point increase in this index implies a 3.8-percentage-point fall in the unofficial economy's share of total GDP. Controlling for

³ In most Asian countries, this index is highly correlated with measures of corruption. Thus, Hong Kong and Korea score 7.5 on this Fraser Institute measure, while Thailand, Malaysia, and Indonesia score 2.5. Singapore is again an anomaly because it scores 0 on this measure, despite having very little corruption.

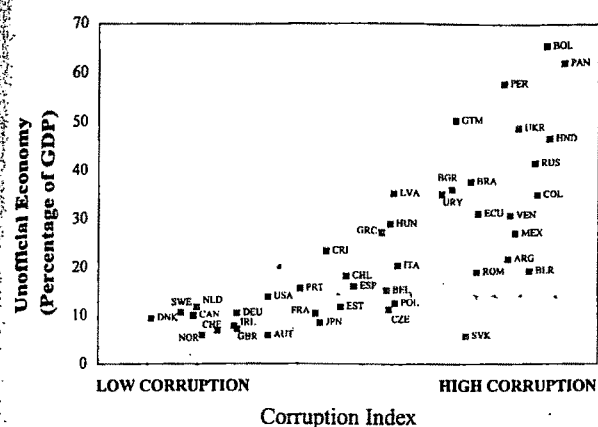


FIGURE 2. UNOFFICIAL ECONOMY AND CORRUPTION

Notes: Unofficial-economy estimates are from Johnson et al. (1998); the corruption index is from Lambsdorff (1998).

log GDP per capita reduces the coefficient to -2.3, but it remains significant.

The extended Transparency International measure of corruption, prepared by Johan G. Lambsdorff (1998), scaled 0–10, covers 43 of the countries in our sample for 1997.⁴ It is higher for countries with less corruption. In our sample, Denmark has the highest score with 9.94 and Bolivia has the lowest in our sample with 2.05. Russia scores 2.27 while Brazil scores 3.56. The best Latin American country is Chile with 6.05. In Table 2 a one-point increase in this index implies a 5.1 percentage point fall in the unofficial economy, and a 3.5-percentage-point fall when the log GDP per capita control is included.

In the Global Competitiveness Survey measure of bribery, scaled 1–7, a higher score means less corruption in 1997 (World Economic Forum, 1997). Among countries for which we also have data on the unofficial economy, the highest score is Sweden with 6.61. The lowest scores (under 3) are for several Central American countries, as well as Russia, which scores 2.72. Brazil scores 3.75. Table 2 shows that a one-point increase in this index implies a reduction in the share of the

unofficial economy by 8.0 percentage points (without the control variable) and by 3.9 percentage points (if we control for log GDP per capita).

In the Impulse index of corruption, a higher score means more corruption (Peter Neumann, 1994).⁵ Russia and Brazil are both awarded 4 out of 5. The best score of 0 is awarded to the usual OECD countries plus Lithuania. As usual, Chile is the best-ranked Latin American country, awarded a score of 1. As Table 2 shows, a one-point increase in this index is associated with a 1.7-percentage-point increase in the share of the unofficial economy. However, controlling for GDP per capita reduces the coefficient by more than half and makes it significant only at the 10-percent level.

In summary, the relationship between share of the unofficial economy and rule of law (including corruption) is strong and consistent across seven different measures. Countries with more corruption have higher shares of the unofficial economy (see Fig. 2). This is true even when we control for income level.

IV. Conclusion

The model of Johnson et al. (1997) has three predictions that find support in the available cross-country data. First, countries with more regulation tend to have a higher share of the unofficial economy in total GDP. Second, a higher tax burden, as perceived by business, leads to more unofficial activity. Third, countries with more corruption tend to have a larger unofficial economy.

This evidence suggests, although it does not prove, that the extent of regulatory and bureaucratic discretion is a key determinant of underground activity. Lax regulations in settings with undisciplined bureaucracies and weak rule of law allow officials to decide individual cases without effective supervision. This creates conditions ripe for corruption (see Kaufmann and Jeffrey Sachs, 1998). Under

⁴ This index requires that countries have had only two (rather than four) surveys. Even in the extended sample, apart from Hong Kong and Singapore, all the other countries that score above 6.5 are long-standing democracies.

⁵ Among the 103 countries surveyed, the worst score is awarded to Bangladesh, Myanmar (Burma), Indonesia, Iran, Nigeria, Pakistan, the Philippines, and Thailand.

such circumstances, many firms choose to operate underground.

REFERENCES

- de Soto, Hernando. *The other path*. New York: Harper and Row, 1989.
- Grossman, Sanford J. and Hart, Oliver D. "The Costs and Benefits of Ownership: A Theory of Vertical and Lateral Integration." *Journal of Political Economy*, August 1986, 94(4), pp. 691-719.
- Gwarney, James and Lawson, Robert, eds. *Economic freedom of the world, 1997 annual report*. Vancouver, BC: Fraser Institute, 1997.
- Johnson, Bryan and Sheehy, Thomas. *Index of economic freedom 1997*. Washington, DC: Heritage Foundation, 1997.
- Johnson, Simon; Kaufmann, Daniel and Shleifer, Andrei. "The Unofficial Economy in Transition." *Brookings Papers on Economic Activity*, Fall 1997, (2), pp. 159-239.
- Johnson, Simon; Kaufmann, Daniel and Zoido-Lobaton, Pablo. "Corruption and the Unofficial Economy." Unpublished manuscript, World Bank, Washington, DC, 1998.
- Kaufmann, Daniel and Sachs, Jeffrey. "Determinants of Corruption." Unpublished manuscript, Harvard University, May 1998.
- Kaufmann, Daniel and Siegelbaum, Paul. "Privatization and Corruption in Transition Economies." *Journal of International Affairs*, Winter 1997, 50(2), pp. 419-58.
- Lambsdorff, Johan G. "Corruption in Comparative Perception," in A. K. Jain, ed., *The economics of corruption*. Boston, MA: Kluwer, 1998 (forthcoming).
- Loayza, Norman V. "The Economics of the Informal Sector: A Simple Model and Some Empirical Evidence from Latin America." *Carnegie-Rochester Conference Series on Public Policy*, December 1996, 45, pp. 129-62.
- Messick, Richard E. *World survey of economic freedom, 1995-1996*. New York: Freedom House, 1996.
- Neumann, Peter. "Böse: Fast Alle Bestechen" ["Flaunting the Rules: Almost Everybody"]. *Impulse*, 4 January 1994, 4, pp. 5-14.
- Political Risk Services. *International country risk guide*. New York: Political Risk Services, 1997.
- Shleifer, Andrei. "Schumpeter Lecture: Government in Transition." *European Economic Review*, April 1997, 41(3-5), pp. 385-410.
- Shleifer, Andrei and Vishny, Robert W. "Corruption." *Quarterly Journal of Economics*, August 1993, 108(3), pp. 599-617.
- . "Politicians and Firms." *Quarterly Journal of Economics*, November 1994, 109(4), pp. 995-1025.
- World Economic Forum. *Executive survey. Global competitiveness report*. Geneva, Switzerland: World Economic Forum, 1997.

Changing Incentives of the Chinese Bureaucracy

By DAVID D. LI*

A striking difference among economies in transition from socialism is in government attitudes toward economic development. In China, the government functions as a "helping hand" for economic development, promoting economic growth; in Russia, the government is like a "grabbing hand," suffocating economic development (Timothy Frye and Andrei Shleifer, 1997). Indeed, after years of transition, it has been increasingly recognized that a proper transformation of the role of government, rather than mechanical implementations of standard reform packages, is a critical determinant of the success of transition.

The Chinese case of transforming government behavior is dramatic and defies conventional explanations. Forty years ago, the same authoritarian regime under the control of the same Communist Party was waging a massive campaign under the name of the Great Leap Forward, resulting in the loss of tens of millions of lives. Thirty-two years ago, the same regime was launching the so-called Cultural Revolution, denouncing any traces of economic incentives. Meanwhile, influential theories of the political economy of the former socialist systems emphasize that unless the one-party (Communist Party) monopoly is abolished, reforms are doomed to fail (Janos Kornai, 1992). These theories cannot explain the Chinese experience, where political liberalization toward representative democracy and the rule of law has been limited and where the one-party monopoly still exists.

This paper argues that, despite the lack of political liberalization, China has benefited

from a major transformation of its bureaucratic system. This transformation, which started years before formal economic reforms, consists of a mandatory retirement program that replaced the revolutionary veterans, a drive for administrative and fiscal decentralization, and the granting of permissions allowing bureaucrats to quit bureaucracy and join businesses. The implementation of these reforms was facilitated by a buyout program for the incumbent government officials. As a result of the transformation, Chinese bureaucrats now face incentives to support economic reform and to promote economic development.

I. Reforms within the Bureaucracy

The first reform, which dealt with the way bureaucrats are promoted and retire, was initiated by Deng Xiaoping in 1980, years before any discussions of reforming the economic system. The proclaimed purpose of Deng's reform was to "abolish the de facto lifetime tenure system of government officials" and to "modernize the contingent of government officials" (Deng Xiaoping, 1983). The reform's crucial measure was to introduce a set of strict retirement ages for government officials and thus, by implication, a massive mandatory retirement program (Hong Yong Lee, 1991; Kenneth Lieberthal, 1995). On a less restrictive basis, an education requirement was also introduced at each level of government positions (see Table 1 for a summary of the reform).

The massive mandatory retirement program was facilitated by a one-time buyout strategy, with the outgoing officials being partially compensated both economically and politically (Susanto Basu and Li, 1996). This is similar in style to the Russian privatization program, in which stakeholders such as managers were offered discounted shares. The buyout program was a special arrangement for revolutionary veterans who were the first and

* Department of Economics, University of Michigan, Ann Arbor, MI 48109-1220. I thank Yasheng Huang, Ramon Myers, Yingyi Qian, Andrei Shleifer, and Wenfang Tang for helpful discussions and detailed comments on earlier versions of the paper. A National Fellowship of the Hoover Institution is gratefully acknowledged. All errors and misinformation are my own.

TABLE 1—THE REFORM OF THE BUREAUCRACY:
FEBRUARY 1982–SEPTEMBER 1984

Statistic	Provincial governors	Ministers	City mayors or department chiefs	County sheriffs or division chiefs
Mandatory retirement age	65	65	60	55
Average retirement age				
Before reform	62	64	58	—
After reform	55	58	50	<45
Percentage with college degree				
Before reform	20	37	14	11
After reform	43	52	44	45
Average tenure (years)				
Pre-1982	6.43/6.23 ^a	6.56	—	—
Post-1982	3.84/4.05 ^a	4.44	—	—

Notes: By 1988, 90 percent of government officials above the county level were newly appointed after 1982; 60 percent of those government officials had college degrees. This was a result of retiring 3.4 million revolutionary veterans.

Source: *Chronicles of Contemporary Chinese Politics* (1996) and Yasheng Huang (1997) (for the tenure data).

^a Governor/party secretary.

biggest potential opponents of the reform. In fact, a special name was coined for this kind of retirement, *lixiu*, meaning literally leaving the post and resting. After *lixiu*, retired officials continued to enjoy all their former political privileges, such as reading government circulars of the same confidentiality level. Some served as special counselors for their successors. As economic compensations, they could keep using their official cars with chauffeurs and security guards. In addition, officials under *lixiu* received an extra month of wages each year and extra housing that their children and grandchildren were entitled to enjoy after their death. Finally, for the most senior officials, there were implicit and informal arrangements whereby their children were allowed to enter politics in senior positions, the origin of the so-called “princeling party” in China.

There have been two major consequences of the reform. The first and direct consequence is that many younger and more educated bureaucrats have replaced the older revolutionary veterans. The new and young officials were generally more supportive of reforms, more adaptable, and more pragmatic. Being better educated in almost all cases, they were also generally more competent than their predecessors. In short, the massive retirement program

has radically changed the human capital of the Chinese bureaucracy. In fact, by 1988, 90 percent of the officials above the county level had been newly appointed after 1982. In contrast, in Russia, Andrei Shleifer (1997) reports that the local leaders are largely the very same people who were there before the reform began.

The second consequence of the reform is that the average duration of a bureaucrat's tenure in a government position has been reduced, partly due to the increased turnover rates of bureaucrats. Compared with the old bureaucracy, the new system, with rather strict age and education requirements, generates more vacancies per unit of time, providing the young and educated with more upward mobility. In addition, the central government seems to have increased the frequency of shuffling provincial/ministerial officials by transferring provincial governors/ministers from one province/ministry to another. Huang (1997) argues that this is due to the central government's fear of local entrenchment due to the decentralization program, which I discuss below. Both effects reduce the duration of bureaucratic tenure in a given position.

Undoubtedly, the changed human-capital composition of the Chinese bureaucracy vastly helped the implementation of the economic reform. This has proved to be a critical factor for the success of reform in many transition economies. In both China and Russia, it has been found that replacing the old management is a critical factor explaining an enterprise's efficiency improvement (Theodore Groves et al., 1995; Nicholas Barberis et al., 1996). Using a large-scale survey, Wenfang Tang and William Parish (1998 Ch. 7) find that the young and educated officials in China are significantly more supportive of reform than are the old and uneducated.

The second reform within the bureaucratic structure is an extensive administrative and fiscal decentralization. During the early 1980's, a so-called fiscal-contract system was introduced for provincial governments in China, in which each province was responsible for collecting tax revenues in its region and was entitled to retain a high proportion of the marginal tax revenue (Jean Oi, 1992; Yingyi Qian and Barry Weingast, 1996). Administrative decentralization was also implemented,

shifting formal authorities from the central government to lower-level governments, including the authority to appoint subordinate government officials (Huang, 1996) and the rights to supervise state-owned enterprises (SOE's). Many provinces, in turn, implemented similar reforms with their subordinate cities or counties.

The extensive administrative and fiscal decentralization has had far-reaching implications for both the Chinese bureaucracy and the Chinese economy. Perhaps the most dramatic effect has been a massive entry of new business entities, which are either partially owned or supported enthusiastically by various governments, whose motivation for doing so comes from increasing tax revenues and, oftentimes, enlarging local employment (Chun Chang and Yijiang Wang, 1994; Jiahua Che and Qian, 1997). Organizationally, these newly established businesses are completely different from SOE's. Typically, the manager is one of the founding entrepreneurs who have contributed their own financial as well as human capital. Often, one of the founding entrepreneurs is a former government official, a phenomenon discussed below. Most important, the manager shares substantial residual control rights with the sponsoring government (Li, 1996). In effect, such new businesses constitute coalitions between the bureaucrats, who contribute bureaucratic connections, and entrepreneurs, who provide business vision and managerial skills. Both parties are indispensable for the success of new businesses in the half-reformed Chinese economy.¹

The massive entry of semi-private businesses has had substantial impact on the incentives and behavior of Chinese bureaucrats. First, bureaucrats are now beginning to act like businessmen. Oi (1992) calls this local corporatism. Also, since bureaucrats now benefit from the new businesses through better official cars, bigger budgets, and better office facilities, their interest is no longer solely in pleasing a bureaucratic superior. More important,

they become de facto shareholders rather than short-termist bribe-takers. This new behavior reduces the prospect for local officials to adopt irrational decisions from the top to maximize their political goods, a bureaucratic pattern associated with the Great Leap Forward.

Second, through their involvement in these businesses, bureaucrats are beginning to become pro-reform in principle, having incentives to lobby higher-level agencies for reduced regulations on behalf of local entrepreneurs. However, one negative effect of having bureaucrats sponsor businesses is that it increases their appetite for bureaucratic power. In the Chinese context, the economic and bureaucratic competition among regions seems to have kept this tendency in check.

II. Changes from Outside the Bureaucracy

A veritable bureaucratic revolution has taken place in China since the mid-1980's, when bureaucrats were allowed to quit their government positions in order to join the business community, a phenomenon that later came to be known as *xiaohai* (i.e., "leaping into the sea" or "jumping into the ocean"). Starting in the mid-1980's, many government agencies began establishing business entities, and bureaucrats became managers of these businesses. Gradually, such businesses are gaining independence from the founding government agencies; meanwhile, many relatively independent and semi-private enterprises are being established. In addition, there were substantial efforts to downsize government agencies during the 1980's, generating a large surplus of government officials. Therefore, by the early 1990's, *xiaohai* was in full swing.

It is not difficult to explain the economics of *xiaohai*. Joining the business world, the former bureaucrats obtain much higher economic payoffs as well as personal freedom, despite being exposed to more economic uncertainty. On the other hand, there is high demand for those bureaucrats, since in the half-reformed economy many nonstate enterprises need their knowledge and skills to deal with the remaining government regulations.

Since 1992, *xiaohai* has been an immensely popular phenomenon among Chinese government officials. In a survey conducted around

¹ Although most studies have focused on those new businesses established by local governments, in fact, many central-government agencies have also established their own business ventures.

1992, Ruying Chen (1993) reports that 30 percent of surveyed officials were thinking about "leaping." In another survey of local government officials in 1995 (State Commission of System Reform, 1996), close to 20 percent were planning on *xiahai*. Of those, 35 percent were looking for joint-venture enterprises, 21 percent for private enterprises, and 15 percent for SOE's. Tang and Parish (1998 Ch. 7) find in their large survey that 99 percent of those officials who planned to quit the bureaucracy wanted to join businesses.

A much more difficult and interesting issue is how to evaluate the impact of *xiahai* on the Chinese bureaucracy. Few systematic analyses exist. A common view, based on experiences of market economies, regards *xiahai* as detrimental to economic growth because it creates an environment where the government and the business are not separated. As a result, government-business collusion arises, and corruption prevails. Moreover, anticipating *xiahai*, according to this view, incumbent bureaucrats may have incentives to increase the complexity of economic regulations so as to increase their future value in the business world.

Contrary to this common analysis, *xiahai* seems to have had a fundamentally positive impact on China's reform process, pushing for dynamic changes in the Chinese bureaucracy in two important ways. First, *xiahai* changes the *ex ante* behavior of bureaucrats before leaving the government by making them more interested in local economic growth, especially in the growth of the non-government-controlled sector (i.e., the nonstate sector), since a more prosperous nonstate sector generates more opportunities for the incumbent bureaucrats when leaving their government positions. Moreover, the *ex ante* behavior is also affected through a reputation effect. That is, an incumbent bureaucrat must establish a pro-business and pro-reform reputation to find a good position in the local business community after leaving the government. The best way to enhance this reputation is to promote growth and reform and to nurture personal rapport with local entrepreneurs by helping their businesses to thrive. In the Chinese business community, personal relations, reputation, and trust (*guanxi*, in general) are very important.

Therefore, a renowned anti-reformist bureaucrat will find it impossible to find a good position in the local business community after leaving his government position.

Second, *xiahai* also transforms the *ex post* behavior of the bureaucrats who quit the bureaucracy. After leaving the government, most bureaucrats joined new businesses rather than traditional state enterprises. Therefore, the former bureaucrats now find that many of the bureaucratic regulations that they helped maintain in the bureaucracy are obstacles to their business interests. They are in the position to help get around and lobby for reductions in these bureaucratic regulations. Having the bureaucratic knowledge and skills, they are more effective lobbyists for reforms of these regulations than are outsiders. This seems to be a rather effective mechanism of reducing Chinese bureaucratic institutions during reform.

Finally, it is interesting to note that *xiahai* seems to be a phenomenon unique to China, at least in its impact on the bureaucracy. In Russia, according to Shleifer (1997), with the exception of Moscow, politicians are not accepted by the private sector. When the private sector grows, they lose power. They are much closer to the formerly state enterprises, many of which are subsidy-seekers. In Poland, local politicians seem to be mostly concerned with getting reelected, rather than joining private businesses after leaving their government positions.

III. Concluding Remarks

The paper argues that the Chinese government's newly acquired spirit for economic development during the reform era is a result of a major transformation of the Chinese bureaucracy, rather than an outcome of political liberalization. Of course, the analysis does not imply that political liberalization is not necessary or unhelpful for economic reform. As many have rightly argued, political liberalization has its own intrinsic value and may be the outcome of an increased per capita income.

Using the framework of Douglass North (1990), the transformation has changed the formal and informal rules within the bureaucracy (i.e. the bureaucratic institutions). The

reformed bureaucratic institutions have induced changes in incentives and behavior of the bureaucrats, and this, in turn, has facilitated reforms of the economic institutions which have spurred economic growth.

China's unusual experience of reforming the bureaucracy without explicitly liberalizing the political system may not be generalizable to other transition countries, since it has been shaped by initial conditions. A particularly intriguing and understudied initial condition is the legacy of the Cultural Revolution, which *ex post* not only boosted Deng Xiaoping's credibility and authority as a reformer, but more importantly, also left China with a weakened bureaucracy for easier bureaucratic reforms.

REFERENCES

- Barberis, Nicholas; Boycko, Maxim; Shleifer, Andrei and Tsukanova, Natalia. "How Does Privatization Work? Evidence from the Russian Shops." *Journal of Political Economy*, August 1996, 104(4), pp. 764-90.
- Basu, Susanto and Li, David D. "Corruption and Reform." Mimeo, University of Michigan, 1996.
- Chang, Chun and Wang, Yijiang. "The Nature of the Township-Village Enterprise." *Journal of Comparative Economics*, December 1994, 19(4), pp. 434-52.
- Che, Jiahua and Qian, Yingyi. "Insecure Property Rights and Government Ownership of Firms." Mimeo, Stanford University, 1997.
- Chen, Ruying. *Xiahai kuangchao (The Craze of Xiahai)*. Beijing: Tuangjie, 1993.
- Chronicles of contemporary Chinese politics. Beijing: People's Publisher, 1996.
- Frye, Timothy and Shleifer, Andrei. "The Invisible Hand and the Grabbing Hand." *American Economic Review*, May 1997 (*Papers and Proceedings*), 87(2), pp. 354-58.
- Groves, Theodore; Hong, Yongmiao; McMillan, John and Naughton, Barry. "China's Evolving Managerial Labor Market." *Journal of Political Economy*, August 1995, 103(4), pp. 873-82.
- Huang, Yasheng. *Inflation and investment controls in China: The political economy of central-local relations*. New York: Cambridge University Press, 1996.
- . "The Industrial Organization of Chinese Governments." Mimeo, Harvard University, 1997.
- Kornai, Janos. *The socialist system*. Princeton, NJ: Princeton University Press, 1992.
- Lee, Hong Yong. *From revolutionary cadres to party technocrats in socialist China*. Berkeley, CA: University of California Press, 1991.
- Li, David D. "A Theory of Ambiguous Property Rights: The Case of the Chinese Non-state Sector." *Journal of Comparative Economics*, August 1996, 23(1), pp. 1-19.
- Lieberthal, Kenneth. *Governing China: From revolution through reform*. New York: Norton, 1995.
- North, Douglass. *Institutions, institutional change and economic performance*. New York: Cambridge University Press, 1990.
- Oi, Jean C. "Fiscal Reform and the Economic Foundations of Local Corporatism in China." *World Politics*, October 1992, 45, pp. 99-126.
- Qian, Yingyi and Weingast, Barry R. "China's Transition to Markets: Market Preserving Federalism, Chinese Style." *Journal of Policy Reform*, 1996, 1(1), pp. 149-85.
- Shleifer, Andrei. "Government in Transition." *European Economic Review*, April 1997, 41(3-5), pp. 385-410.
- State Commission of System Reform. "The Environment of Changes in the Role of the Government in China: Analysis of a Survey," in *Chinese Economic Almanac 1996*. Beijing: Statistic Publishing House, 1996, pp. 976-80.
- Tang, Wenfang and Parish, William L. *The changing social contract—Chinese urban life under reform*. New York: Cambridge University Press, 1998.
- Xiaoping, Deng. *Deng Xiaoping Wenxuan [Selected Works of Deng Xiaoping]*. Beijing: People's Publishing, 1983.

Private Enforcement of Public Laws: A Theory of Legal Reform

By JONATHAN R. HAY AND ANDREI SHLEIFER*

In the last several years, the countries of Eastern Europe and the former Soviet Union (FSU) have made tremendous progress in price liberalization, privatization, and macro-economic stabilization—the standard steps of the so-called shock therapy. Yet in the aftermath of these reforms, the East European countries have begun to grow rapidly, while the countries of the FSU, particularly Russia, are at best beginning to turn around. It is not possible to explain these differences in performance in terms of either having too much shock therapy or not enough of it, since the reforms that the different countries have pursued have been broadly similar.

A more plausible reason for the difference in performance is that institutional reforms, such as those of government regulation, the legal system, and the bureaucracy, have advanced much further in Eastern Europe than in Russia (Shleifer, 1997; Simon Johnson et al., 1997). Indeed, institutional failures have arguably deterred small-business formation, foreign investment, and enterprise restructuring in the FSU. For growth to take off, Russia and other FSU countries must radically improve the quality of their institutions.

In this paper, we discuss the principles of perhaps the key institutional reform, that of the legal system. The ideas we describe were developed at the Institute for Law Based Economy in Moscow. Since its inception in 1994, the Institute has been a key player in the Russian legal reform, and we were both involved in its work. Despite the Russian specificity of some of the analysis, we believe that these ideas apply to the problems of legal reform in the rest of the FSU, as well as in emerging economies more generally.

I. The State's Failure To Provide and Enforce Laws

Business people in Russia use the state legal system a lot less than they feel they need to. In a survey of shopkeepers in Russia and Poland, for example, 45 percent of Moscow respondents said that they needed to use the courts in the last two years but did not, compared to only 10 percent of the Warsaw respondents who gave this answer (Timothy Frye and Shleifer, 1997). There are two apparent reasons why the state legal system is not used in Russia: the low quality of the services it provides and the unwillingness of business people to expose themselves to the legal system, and to the government, more generally.

The quality of the legal system is notoriously bad (Avner Greif and Eugene Kandel, 1995; Katharina Pistor, 1996). The legal rules are incomplete in crucial areas needed to support existing business activity, such as real-estate registration. When legal rules do exist, in many instances judges do not know what they are. Many judges, for example, are unfamiliar with the relatively new securities law, which comes up in securities-markets disputes. Even when the law speaks to a particular matter, judges may not have the resources or inclination to verify the relevant facts. And when the facts are available and the legal rules exist, judges may be biased, corrupt, or partial to political sentiment, and hence it is by no means certain how they will rule. Finally, once a judge rules, there are often no institutions to enforce his ruling. For example, contracts between Russian and Western partners often specify London courts as the venue for dispute resolution. When the Russian partner breaks the agreement and the London court rules against him, the Western partner is still left with absolutely no mechanism of collecting his claims.

* Ironwood Holdings and Department of Economics, Harvard University, Cambridge, MA 02138, respectively. We thank Edward Glaeser and Avner Greif for comments. This paper is dedicated to the memory of Albert Sokin.

A further reason that private parties in Russia refuse to use the legal system is that they operate to some extent extralegally to begin with and, hence, do not want to expose themselves to the government. The tax system in Russia is sufficiently arbitrary and draconian that private firms are either in violation of tax law or even operate unofficially. By some calculations (Johnson et al., 1997), over 40 percent of the Russian economy is unofficial. Given that many firms have both some official and some unofficial business, the majority of Russian businesses are probably in violation of some tax, customs, foreign-exchange, or regulatory rules and, hence, would not use the official legal system to resolve disputes for fear of exposure.

The consequence of this avoidance of the legal system is that private rather than state mechanisms are used to resolve disputes. These mechanisms range from social norms and pressures, to arbitration, to employment of private but legal protection agencies, to organized crime. In some cases, where parties interact repeatedly and the stakes in individual disputes are small compared to the value of long-term relationships, peaceful private mechanisms work extremely well (Robert Ellickson, 1991, Lisa Bernstein, 1992, Greif, 1996). For example, arbitration succeeded as a means of resolving disputes between brokers on Russia's commodity and stock exchanges, who interact repeatedly and can use the exchange to enforce the arbitrator's decisions (Frye, 1996). Even illegal private enforcement organizations that gain monopoly in dispute resolution in a particular area and manage to gain acceptance for their rules and enforcement mechanisms may be reasonably efficient. Why, then, has the private resolution of disputes left Russia with what is widely regarded as a dysfunctional legal system?

The trouble is that private dispute resolution often does not work efficiently. Many commercial disputes do not fit the nice picture of repeated interactions over long periods of time, where access to the system is a valuable asset that the trading parties would not give up. This is so with debt collection, where borrowers are too far underwater to worry about the future, or with big ownership disputes. In

these and other cases, there needs to be some force in enforcement.

Moreover, private rules, including those for using force, are often neither known nor accepted by the disputing parties. If person A borrows money from person B and does not repay, A's protectors might think that the appropriate rule is to extend the period of repayment, whereas B's protectors might think the appropriate rule is to kill A. Once A is dead, there may be no public lesson to be learned about what the rules are, since the potential future borrowers from B or other lenders, including A's associates, would not even generally know what rule A has violated and why he was killed. Private rules are often unrecognized, unknown, and not enforced consistently, which makes it prohibitively expensive for private parties to rely on them to structure transactions.

Last but not least, private enforcement is unhelpful in legal disputes with the government. As a consequence, private parties remain vulnerable to the threat of discretionary regulation and extortion by public officials, without any effective legal recourse (see Shleifer and Robert Vishny, 1993). The standard function of the judicial system of providing a check, however rudimentary, on other branches of government is lost with purely private enforcement of private rules.

In sum, private enforcement of private rules in Russia has emerged as a market response to the failure of the state to provide and enforce its own rules, largely because of very weak incentives in the government to provide law and order. This private mechanism has the advantage that both the disputing parties and the enforcers have economic incentives to pursue enforcement. Yet it also has the major disadvantage that private rules are often different for different enforcers, insufficiently well known, and not legitimate enough for business people to rely on them in structuring their transactions. The result is that the legal system is viewed as a failure, and a lot of trade and production simply does not take place.

A common recommendation to address this problem, in line with the traditional economists' view that laws should be publicly enforced (Douglass North, 1981), is to beef up

the state legal system, through administrative reforms of police and judicial system, accelerated production of laws, training of judges, and so on. Unfortunately, such recommendations often overlook the fundamental problem that the incentives in the government to reform itself are lacking, and hence these reforms may fail or even backfire. When the elite units of the Russian police obtained bigger guns to fight the mafia, they simply sold these guns to the mafia at higher prices than the previous, less powerful, weapons could fetch. The reforms of the tax bureaucracy have not gone well either, in part because no government official has enough authority to shake up a system that benefits, at least indirectly, other government officials. And increases in the power of tax police have led to greater arbitrariness, abuse, and corruption. Without a "benevolent" dictatorship, a common collective memory of law and order, or at least a very strong and unified democratically elected government (none of which describes the reality of Russia at the moment), strengthening the state's legal apparatus can do more harm than good. What, then, can be done in the interim to improve law and order?

II. Private Enforcement of Public Rules

The principal argument of this paper is that the appropriate legal-reform strategy for a country like Russia is private enforcement of public rules. Public rules can address the problems of multiplicity, obscurity, and illegitimacy that plague the private rules. Private enforcement of these rules introduces powerful incentives that the public sector does not have.

The strategy of private enforcement of public rules begins with the creation of legal rules that can be enforced jointly by an extremely limited public judicial system and the much more extensive private enforcement system. From the viewpoint of the state legal system, public laws can help well-intentioned judges to resolve disputes, and even restrict the discretion of the not-so-well-intentioned judges. To the extent that laws make judges more predictable, business will rely on these laws more and hence demand the services of the official legal system.

Even when disputes are ultimately resolved by courts, however, much of the benefit of public rules comes from private parties structuring their transactions so that courts become more usable. For instance, suppose that a rule completely prohibits, and actually annuls, certain "self-dealing" transactions by corporate managers, such as sales of corporate assets to affiliated parties. When this rule exists, large shareholders will try to institute procedures that allow them to review corporate transactions and to document who the buyers are. Most of the information collection will be done by private parties who have powerful incentives to verify violations, and who can then come to a court for a very simple decision based on verifiable information that the private parties themselves provide. All a judge has to do is annul the sale. The likelihood of such annulment would make managers wary of breaking the law, and buyers wary of losing their money. Moreover, when public rules exist and are clear, judges would come under public pressure to enforce them, rather than rule corruptly, politically, or arbitrarily. Without a specific rule, it is not clear what large shareholders need to do or to document in order to use the court. Public rules work because they tell private parties what they need to do to use the legal system and thus provide incentives for them to use it.

But public laws have a further, perhaps even more significant, benefit in an emerging economy: they become the focal point of totally private contract enforcement and dispute resolution. Unlike the private rules, public laws are public, and hence private enforcers can free ride on them to structure their own activities, and to create their own reputations. Public rules can thus coordinate the expectations of market participants even with little public enforcement, similarly to the idea of coordination of beliefs in Thomas Schelling (1960), Robert Sugden (1989), and Greif (1994). Public law is particularly attractive for belief coordination because in the eyes of many people law has a degree of legitimacy that private rules do not have. In an emerging economy, these coordination benefits of public rules may be enormous.

Take the case of reputation development by private enforcers. A public rule on loan de-

faults may encourage both A's (the borrower's) and B's (the lender's) protectors to use it to resolve the dispute. Suppose this rule is quite unfavorable to the lender, B. B's protectors may still accept it, because they can then become known for enforcing widely accepted public laws and thus further their public reputation. Indeed, it may no longer be in the interest of either A's or B's protectors to enforce their own rules, because these rules would be much less well understood, undermining their reputations, and hence the demand for their services. Indeed, if B's protectors try to enforce some other rules, A's protectors can tell all market participants that B's protectors are acting arbitrarily, thereby triggering some form of collective punishment or exclusion. Last, but not least, other borrowers (and lenders) can now structure their contracts with better knowledge of what happens when a borrower defaults. A's and B's protectors now have a larger share of a larger market.

Interestingly, private enforcement organizations in Russia often ask disputing private parties for copies of their written agreements. The enforcers do not want to enforce arbitrary claims; they want to establish reputations for resolving disputes according to rules, and public rules are a chosen focal point. As a further step, dispute resolution can proceed in the shadow of private law-enforcement organizations, but without their direct participation.

Through these mechanisms, public rules acquire a reputation and legitimacy of their own. In some cases, these rules are enforced by courts, though with significant efforts by private parties to simplify the courts' decision process. In other cases, these rules are enforced by private parties without any reliance on courts. In still other cases, the parties to a dispute agree to a resolution in line with these rules without any help with enforcement, since they know what is going to be enforced. In all these ways, private enforcement of public rules can work reasonably efficiently even when public enforcement remains ineffective.

III. Which Public Rules Encourage Private Enforcement?

So far, we have spoken of legal rules, and their coordination benefits, in general terms,

without distinguishing between good and bad rules. But, of course, good rules are more likely to be used by economic agents, as well as by both the private and the public enforcers, than bad rules. What, then, constitutes good legal rules?

In line with Section II, good legal rules are those likely to be adopted by private parties for both structuring and enforcing their transactions, as well as used by courts. In general, rules that are usable by private parties and those usable by courts are the same, since both types of users are looking for simplicity, consistency with standard business practice, efficiency, the ease of verification of violations, and most importantly, effectiveness of enforcing decisions.

The standard rule-making strategy is to borrow legal rules from advanced countries, rather than to reinvent the wheel. Indeed, virtually every country in the world has borrowed most of its commercial law from a few legal systems, particularly French and German civil law and English common law (Alan Watson, 1974; René David and John Brierley, 1985). But the decision to borrow does not end the story. Legal rules both within and across traditions vary enormously, and some rules facilitate trade better than others (Rafael La Porta et al., 1998). There is thus a question of which rules to pick. More importantly, rules are specific to other elements of the legal system. In particular, Western legal rules rely on vastly more extensive judicial verification and public enforcement mechanisms than are available in a transition economy. Western rules must therefore be adjusted to facilitate private enforcement.

On this basis, we suggest three general lessons for developing legal rules for a country like Russia. First, as argued by Bernard Black et al. (1996) and Hay et al. (1996), it is better to have "bright line" rules (i.e., rules that make it easy for judges, or private enforcers, to verify violations). For example, in an advanced economy, it may be best to have a flexible anti-self-dealing rule that allows managers to undertake transactions as long as (they can show in court) these are in the interest of shareholders. Such a rule would not work in Russia because a judge could not use it and is likely to side with the manager, or with

whoever pays him more, in a dispute. A better rule for Russia would prohibit and annul all transactions between the corporation and any entity in which the manager has an interest. Violations of this rule are easier to verify and punish, even if it is less flexible.

A second idea, which has not had sufficient impact on Russian law-making, is that there needs to be a private right of action, and a clear private remedy, in a dispute. In the previous example, if shareholders have no clear right to sue for self-dealing and be rewarded through a higher value of their shares (or even a part of a fine), the likelihood of enforcement is negligible. Counting on an administrative agency to enforce the law is a mistake, since the agency is more likely to listen to corporate managers than to complaining shareholders. If, in contrast, a large shareholder can bring a specific violation to a court, he is more likely to monitor the managers. The role of private enforcement in making this law work is overwhelming.

Third, whenever possible, laws must agree with prevailing practice or custom. If public laws violate the practice, then private parties may refuse to enforce them either on their own or with ultimate reference to courts. The coordination benefit of public laws would then be lost. Alternatively, when laws absolutely must change the existing practice, it is crucial to write them keeping in mind what groups of private agents would enforce them. Thus the Russian mass privatization program relied on the incentives of corporate managers to implement its rules (Maxim Boycko et al., 1995), whereas the fledgling Russian corporate law relies crucially on the incentives of large private shareholders to control the managers.

To summarize, it is difficult but possible to construct legal rules, based on adjusting the best world practice, that become the focal point of both private and public enforcement, and that are usable even in a country with extremely limited public enforcement of laws, largely because they make private parties do most of the enforcing.

IV. Toward Public Enforcement

Although this paper has advocated the benefits of private enforcement of public laws, ultimately, as a country develops, the role of the

public sector in law enforcement is likely to increase. The final question we address is what can be done to improve public law enforcement.

In many cases, public law enforcement is most likely to benefit from institutional reforms outside of the law-enforcement sector proper. In Russia, these reforms include tax reform and federalism reform (see Johnson et al., 1997; Shleifer, 1997). The simplification of tax rules, combined with the reduction of marginal rates, would draw firms out of the unofficial economy thus increasing the demand for official law enforcement and reducing the demand for unofficial services. Federalism reform can create the incentives for regional governments to provide high-quality courts to attract business and expand the tax base, as well as to improve police and other protective services for business. Indeed, these incentive-based reforms are likely to do more for law enforcement than the difficult-to-implement administrative reforms, which, as we mentioned earlier, can fail or even backfire.

Public enforcement is surely the ultimate goal of any legal reform. Yet it is important to remember that the strategy of private enforcement of public rules can serve Russia, and many other emerging economies, extremely well in the short and medium term.

REFERENCES

- Bernstein, Lisa. "Opting Out of the Legal System: Extralegal Contractual Relations in the Diamond Industry." *Journal of Legal Studies*, January 1992, 21(1), pp. 115-57.
- Black, Bernard; Kraakman, Reinier and Hay, Jonathan. "Corporate Law from Scratch," in Roman Frydman, Cheryl Gray, and Andrzej Rapaczynski, eds., *Corporate governance in Central Europe and Russia*, Vol. 2. Budapest, Hungary: Central European University Press, 1996, pp. 245-302.
- Boycko, Maxim; Shleifer Andrei and Vishny, Robert W. *Privatizing Russia*. Cambridge, MA: MIT Press, 1995.
- David, René and Brierley, John. *Major legal systems in the world today*. London: Stevens, 1985.

- Ellickson, Robert C. *Order without law*. Cambridge, MA: Harvard University Press, 1991.
- Frye, Timothy. "Contracting in the Shadow of the State: Private Arbitration Courts in Russia." Mimeo, Harvard University, 1996.
- Frye, Timothy and Shleifer, Andrei. "The Invisible Hand and the Grabbing Hand." *American Economic Review*, May 1997 (*Papers and Proceedings*), 87(2), pp. 354-58.
- Greif, Avner. "Cultural Beliefs and the Organization of Society: A Historical and Theoretical Reflection on Collectivist and Individualist Societies." *Journal of Political Economy*, October 1994, 102(5), pp. 912-50.
- . "Contracting, Enforcement, and Efficiency: Economics Beyond the Law," in Michael Bruno and Borislav Pleskovic, eds., *World Bank Annual Bank Conference on Development Economics*. Washington, DC: World Bank, 1996, pp. 239-65.
- Greif, Avner and Kandel, Eugene. "Contract Enforcement Institutions: Historical Perspective and Current Status in Russia," in Edward P. Lazear, ed., *Economic transition in Eastern Europe and Russia: Realities of reform*. Stanford, CA: Hoover Institution Press, 1995, pp. 291-321.
- Hay, Jonathan; Shleifer, Andrei and Vishny, Robert W. "Toward a Theory of Legal Reform." *European Economic Review*, April 1996, 40(3-5), pp. 559-67.
- Johnson, Simon; Kaufmann, Daniel and Shleifer, Andrei. "The Unofficial Economy in Transition." *Brookings Papers on Economic Activity*, 1997, (2), pp. 159-239.
- La Porta, Rafael; Lopez-de-Silanes, Florencio; Shleifer, Andrei and Vishny, Robert W. "Law and Finance." *Journal of Political Economy*, 1998 (forthcoming).
- North, Douglass C. *Structure and change in economic history*. New York: Norton, 1981.
- Pistor, Katharina. "Supply and Demand for Contract Enforcement in Russia: Courts, Arbitration, and Private Enforcement." *Review of Central and East European Law*, 1996, 22(1), pp. 55-87.
- Schelling, Thomas. *The strategy of conflict*. Cambridge, MA: Harvard University Press, 1960.
- Shleifer, Andrei. "Government in Transition." *European Economic Review*, April 1997, 41(3-5), pp. 385-410.
- Shleifer, Andrei and Vishny, Robert W. "Corruption." *Quarterly Journal of Economics*, August 1993, 108(3), pp. 599-618.
- Sugden, Robert. "Spontaneous Order." *Journal of Economic Perspectives*, Fall 1989, 3(4), pp. 85-97.
- Watson, Alan. *Legal transplants*. Charlottesville, VA: University of Virginia Press, 1974.

FORECASTING JAPAN'S FUTURE: THE LESSONS OF HISTORY[†]

The 1940 System: Japan under the Wartime Economy

By YUKIO NOGUCHI*

Many observers point out that the economic system in Japan is fairly different from those of other capitalistic countries. The basic elements which constitute the so called "Japanese economic system" can be identified as follows: (i) employment practices such as the lifetime-employment, the seniority wage, and in-house labor unions, which are typically observed in large organizations; (ii) weak influence of shareholders on corporate decisions; (iii) an indirect financial system; (iv) intensive government interventions on private economic activities, in particular, protection of low-productivity sectors such as agriculture and small businesses; (v) a tax structure which relies heavily on income tax, and a government structure in which local governments are strictly controlled by the central government.

Many argue that the above features are reflections of the unique aspects of Japan's cultural and social norms whose origins can be traced back to the history of Japan. I will argue that these are not necessarily "intrinsic Japanese" and that most of them were introduced as the wartime system during the years around 1940, and hence can be called the "1940 system" (Noguchi, 1995).

I. The Origin of the "Japanese Economic System"

One of the most important features of Japanese corporations is that their main objective is to advance interests of their employees rather than to maximize profits. In other words, Japanese corporations are cooperative

organizations of workers and managers. They work together for a single common goal: the growth of their corporation. Many analysts argue that this is the result of the group-oriented nature of the Japanese society.

It must be noted, however, that in prewar Japan, chief executives of corporations were mostly major shareholders. Because of their dominance in decision-making, corporations were managed so as to maximize profits for the sake of shareholders. In this sense, Japanese corporations were not significantly different from those in other capitalistic countries. This corporate structure was changed drastically in the process of preparation for the war. Under the 1938 National Mobilization Act (*Kokka Sodo-in-ho*), limitations on dividends were imposed, shareholders' rights were limited, and corporations were restructured so as to serve the collective interests of employees. These measures were taken in order to promote workers' incentives and to raise the productivity of manufacturing industries.

It was also during this period that such practices as lifetime employment and the seniority wage which had been introduced in large corporations after World War I became prevalent. This change was accelerated by the Wage Control Ordinance and other government initiatives. The nature of labor unions also changed significantly under the wartime regime. Industry-based unions were dissolved and replaced under government initiatives by corporate in-house unions called *Sangyo Hokokukai*. This can be regarded as the origin of the present Japanese labor unions. Subcontracting is another feature of Japanese industry. The origin of this system can also be found in the emergency measures introduced in the wartime to increase the production of munitions. More than 40 percent of subcontracts to

[†] *Discussant*: Hugh Patrick, Columbia University.

* Research Center for Advanced Science and Technology, University of Tokyo, 4-6-1 Komaba, Meguro-ku Tokyo 153, Japan.

supply parts for Toyota Motor Company in the 1960's were formed in the wartime.

It was also during the war that many of the representative Japanese companies grew to the present giant status (John W. Dower, 1993). This is particularly apparent in the automobile industry. Before the war, Japan's automobile industry was very weak. The production of passenger cars was dominated by Ford and General Motors, both of whom had their own assembly plants in Japan. It was only after the Manchurian Incident of 1931 that situations began to change. Toyota Automatic Loom started building cars, and Kaishinsha and Ishikawajima-Harima Heavy Industries formed a joint venture called Nissan. Five years later, the Automobile Manufacturing Business Act (*Jidosha Seizo Jigyo-ho*) was passed. Its aim was to "establish the automobile industry in order to contribute to national defense and the industrial development of Japan." Under this law, licensed Japanese carmakers were to be given preferential treatment as regards taxation and financing. To protect the domestic producers further, tariffs were raised, and restrictions were placed on imports. This legislation intended to drive Ford and General Motors out of the Japanese market. The first to benefit from these measures were Toyota and Nissan. The legislation clearly helped Japan's automobile industry get off the ground. Similar patterns are observed in the electric machinery industry. Both Hitachi and Toshiba, Japan's largest manufacturers of electrical equipment, were established in the early 1900's, but it was in the late 1930's that they emerged as comprehensive producers of electrical machinery.

Similar trends can be observed in the press. The three major national dailies Yomiuri, Asahi, and Mainichi each trace their beginnings to the 19th century, but not until the war were they able to substantially increase both circulation and influence. In 1941, the Japanese government enacted the Newspaper Business Ordinance, under the auspices of the National Mobilization Law. One year later, the Japan Newspaper Association was established. These steps paved the way for a series of consolidations. Together with the "one-daily-per-prefecture" policy advocated by the Ministry of Interior, this drive toward consolidation

dramatically reduced the number of daily newspapers: from 848 in 1939 to only 54 in 1942. The Japanese dailies available today are in fact the product of the wartime consolidation.

Until the 1930's, capital markets were quite active in Japan, and many major corporations obtained funds directly from those markets. This was because few regulations existed in the financial markets. In 1931, 87 percent of new funds were supplied from the capital markets, and bank lending played only a minor role. Reflecting this financial structure, shareholders' influence on corporate decisions was strong, as mentioned above.

Regulations on financial industry were strengthened during the late 1930's. Strict limits were imposed on dividends, and as a result, fund-raising from the stock market became difficult. On the other hand, the government took initiatives to increase the supply of funds from banks, in particular, government-supported long-term credit banks, the most notable of these being the Industrial Bank of Japan (*Nihon Kogyo Ginko*). On the other hand, consolidations of commercial banks were accelerated by government initiatives, and their number was reduced from 1,402 in 1926 to 186 in 1941, and further to 61 in 1945.

The origins of the present Japanese government structures can also be found in the wartime. The present Japanese tax system, which relies heavily on the income tax and the corporate tax, is very different from that in the prewar era in which traditional taxes such as land tax, liquor tax, sugar tax, and customs duty contributed to national revenues. It was the wartime reform that brought fundamental changes in the tax system. In the tax reform of 1940, a source withholding tax on wage income was introduced, and the corporate income tax was newly established. This reform has enabled the government to tax the modern sector of the economy. The revenue was used not only for military purposes, but also to help the economically distressed rural areas. In order to facilitate income transfers from the modern sector of the economy to rural areas, a subsidy system from the central to the local governments was established in the 1940 reform. This changed the relationship between the central

and the local governments, from the prewar system in which local governments' autonomies were strong, to the present system in which local governments are subordinates of the central government.

Similar changes are observed in the government-business relations. In prewar Japan, private enterprises were to a large extent independent from bureaucratic interventions. During the wartime, interventions in private business activities increased. Also, such regulations as the Foodstuff Control Act and the Land and House Lease Act were either newly introduced or strengthened during this period.

The above discussion suggests that the prewar Japanese economic system was in general more market-oriented, and in that sense more similar to the "Western system" than is the present one. Introduction of the total war system changed the Japanese economic system. In terms of continuity versus discontinuity, it can be said that a discontinuity can be found around the year 1940. This is different from the conventional view that the Japanese system is an indigenous system and, hence, cannot be changed easily. The implication of the argument presented here would be that the Japanese system can in principle be changed because it is an "artificially introduced" system.

II. The 1940 System and Rapid Economic Growth

Another topic concerning continuity versus discontinuity is whether a discontinuity can be found at the end of the war. It is generally believed that, with the end of World War II, Japan was reborn because it was totally reformed by the postwar democratization policies. For decades, this has been the orthodox interpretation of the postwar history of Japan. The view presented here is radically different because it argues that the present system is a remnant of the war system.

It is true that the end of the war meant a transformation in certain aspects of Japanese society. In fact, the country changed from a military state to one committed to peace. The drawing up of a new Constitution and the purge of 1946 were followed by a series of reforms. These included the dissolution of the large financial conglomerates (*zaibatsu*), the ag-

ricultural land reform, and the enactment of labor laws. It is believed that these postwar reforms were the fundamental causes that made Japan's postwar economic growth possible.

However, the wartime system survived. If one goes below the surface, one finds that a number of wartime systems and mechanisms still form the basic structure of the present Japan's economy. Indeed, the wartime legacy is so prevalent that the current economic system should be regarded as something that has its roots in the wartime as opposed to being postwar.

At the end of the war, the military was completely dismantled by the Allied Forces. However, since the Allied Forces chose the indirect occupation policy, the bureaucratic system was left intact, except for the Ministry of Interior which was split into several ministries such as the Construction Ministry and the Local Autonomy Agency. As a result, a remarkable continuity in the bureaucratic system was brought about. The continuity can be seen in the number of purged bureaucrats. Out of the 21,000 purged people, only 2,000 were former bureaucrats, and most of them were from the Ministry of Interior. In the case of the Ministry of Finance, the number of purged people was only nine.

The tax system which was introduced by the 1940 tax reform remained. And most other governmental structures remained. For example, despite the call for local autonomy, the subordinate status of local governments did not change, because revenue sources continued to be grasped by the central government.

One of the most symbolic evidence of the continuity would be the (old) Bank of Japan Law, which remained until 1997 as one of the most important laws guiding Japan's financial system. The law, modeled on the German Reichsbank Law of 1939, was enacted in 1942 in order to put the finishing touches to the wartime financial controls. It declared that "The Bank of Japan must perform its mission solely for the purpose of accomplishing the goals of the nation."

The relics of the wartime regime are not fragmentary or exceptional in the financial system. With the end of the war, the special banks abroad were closed, and two domestic special banks were converted to commercial banks. On the other hand, such new institu-

tions as the long-term Credit Bank of Japan (*Nihon Choki Shinyo Ginko*) and the Development Bank of Japan (*Nihon Kaihatsu Ginko*) were created as successors. These special banks continued to dominate the nationwide flow of funds until the mid-1950's. Similar continuity can be observed for commercial banks. As mentioned before, the number of commercial banks was reduced to 61 by the wartime control. There has been little change in the number since then. Given the extraordinary growth of the economy, it is surprising that the structure of the banking industry changed so little.

As seen above, the shadow cast by the relics of the wartime system is far-reaching; indeed, it can be said that the wartime regime was the most important driving force behind the rapid economic growth of postwar Japan. First, the group-oriented nature of the Japanese corporations enhanced workers' morale, and in-house labor unions did not oppose the introduction of new technologies. Second, the indirect financing system contributed to concentrating economic resources to strategic industries. Without such financing, resources would necessarily have been diverted during the early 1950's to projects that yield quick returns and, as a result, the heavy-industry-led economic growth would not have materialized. Third, government protections on low-productivity sectors mitigated social frictions that would have been brought about by rapid growth of the leading industries.

The system also worked very efficiently in the process of adapting to the oil crisis during the 1970's, because in-house labor unions did not demand radical wage increases. Workers knew that they and corporate managers were in the same boat, so that excessive wage demands would sink the boat and drown everyone on board.

In this way, a system originally designed to fight the total war continued to play an integral role in the management of the Japanese economy. The aim of mobilizing the total power of the national economy has, of course, changed from military to economic, but the all-out nature of the system itself has remained unchanged.

III. The Need to Overcome the 1940 System

The surrounding conditions have changed, however, in recent years. The Japanese economy now faces a need to fundamentally change its industrial structure in order to adapt to the new economic environments. The essential difference from the rapid-growth era is that future courses must be found by a trial-and-error process. Under such conditions, individual creativity rather than group cooperation is required. The 1940 system works well in an environment in which members must cooperate for a single and well-defined goal, but it does not work well under conditions in which change is needed.

The system has become an obstacle, particularly in the following respects. First, it is difficult for workers to move from one company to another under the lifetime employment and seniority wage system. This works as an obstacle to changing the industrial structure. Second, under the indirect financial system, it is difficult to supply investment funding for venture businesses whose futures are uncertain. Although several attempts have been made to finance venture businesses, only insufficient funds have been provided for them.

It is therefore quite important to overcome the 1940 system for the future of the Japanese economy. This is far from an easy task, however, because the system covers almost every area of the economy. On the other hand, it is in principle possible to overhaul the system, because the 1940 system is a system that has been "artificially" introduced. It may be that Japan's future is considerably endangered because of the 1940 system, or it may be that the changing economic environments and the development of new technologies will destroy the system. In this sense, Japan is at a very important branching point of its history.

REFERENCES

- Dower, John W. "The Useful War," in John W. Dower, *Japan in war and peace*. New York: New Press, 1993, pp: 9-32.
- Noguchi, Yukio. *1940 Nen-taisei [The 1940 system]*. Tokyo: Toyo Keizai Shimposha, 1995 (in Japanese).

Structural Change and Japanese Economic History: Will the 21st Century Be Different?

By GARY R. SAXONHOUSE*

There is very considerable disagreement among economists as to how much is known about the process of economic development and why it is that some nations grow faster than others (N. Gregory Mankiw, 1995). This is reflected in the complementary disagreement among economists as to what policies ought to be recommended to governments eager to enhance their economic performance. Should governments be preoccupied with the proportion of gross domestic product that is devoted to savings and investment and with the rate of population growth; or should priority be given to the provision of public goods? Is technology a global public good, or should governments be preoccupied with how to enhance technological transfer and with international economic relations? Are macroeconomic issues really central to an understanding of economic growth and performance? In a world of imperfect information, should not the focus of attention be on the design of efficient systems and monitoring relationships and on how best to assign property rights?

These uncertainties in the scholarship on economic growth complicate any assessment of Japan's experience. It is also fair to say that were Japan's remarkable economic performance better understood the study of economic growth would today be on much firmer foundations. It is clear that some time after 1913, Japan's rate of economic growth ratcheted upward relative to what are now thought of as the advanced industrialized economies. Thereafter, wartime and its aftermath apart, until the early 1990's, Japan's rate of economic growth was persistently higher than those economies, but by more or less the same degree. Japan's per capita rate of economic growth trebled in 20 years (1953–1973) after

wartime recovery by comparison with the 25 years (1913–1938) before the war, but this was also true on average of the other advanced industrialized economies. In the two decades after 1973, Japan's per capita rate of economic growth fell by three-fifths, but once again this was also true, on average, for the other advanced industrialized economies.

I. Japan's Higher Growth Path

Japan's record raises a number of questions. What is it that put Japan on this high-growth trajectory? Is Japan's 20th-century growth experience best understood as convergence toward a steady-state rate of growth? If so, are there processes endogenous to Japan's growth performance that have altered the speed of Japan's convergence to its steady-state growth rate or indeed, Japan's steady-state growth rate itself? How have Japan's distinctive but malleable economic institutions interacted with Japan's growth process? In the perspective of Japan's 20th-century growth experience, what will be Japan's future growth path? Are the 1990's to be thought of as a cyclical downturn for Japan, something akin to the cyclical downturn of the 1920's which followed the long-ago bubble years of World War I? Will Japan reemerge from this downturn and once again outperform the other advanced industrialized economies as it has for most of the 20th century?

It is plausible that the Japanese government policies that pushed the rapid accumulation of human capital in the late 19th century and early 20th century helped create the conditions that subsequently put Japan on a relatively high-growth path. By the second decade of the 20th century, Japanese workers already had two-thirds as many years of formal education as American workers even while working with no more than $\frac{1}{15}$ of the physical capital. Compared to its level of per capita gross domestic product, human capital was abundant.

* Department of Economics, University of Michigan, Ann Arbor, Michigan 48109.

The impact of this growing accumulation of human capital was substantial. In the mid-Meiji period, Japanese with advanced formal education were absorbed as teachers and administrators in a rapidly expanding educational system and in newly created positions within the Meiji government. By the time of World War I, however, Japanese with advanced formal training in science and engineering came to be employed in manufacturing industries in large numbers, having a seemingly significant direct impact on productivity (Saxonhouse, 1977). More human capital raised the ability of these industries to absorb new technologies, lowering the cost of imitating ideas and approaches that were developed both outside and within Japan. Note that it was only after Japan had accumulated substantial human capital that it was able to experience relatively rapid total factor productivity growth. Earlier, between 1890 and 1913, not unlike Singapore between 1966 and 1990, Japan experienced a substantial decline in total factor productivity (Angus Maddison, 1995).

II. Japan's Institutional Adaptation

Japan's superior growth performance has continued for much of the 20th century even as its economic institutions have changed markedly. In the early decades of the 20th century, rapid economic growth in Japan coexisted with very high labor turnover and the elaborate monitoring of production workers even while widespread use was made of piece-rate systems of compensation. Just a few decades later, permanent employment came to be characteristic of most large Japanese firms with pay closely related to the number of years of employment. The market for experienced managers, engineers, and workers that had been active earlier in the century withered.

Dramatic change was not confined to Japan's labor markets. The Japanese economy has grown relatively rapidly both with rather unregulated and with heavily regulated financial systems. During the early decades of the 20th century, there was virtually free entry in Japan's banking sector. Despite the presence of a small number of banks that were affiliated with the elaborate cross-holding networks known as *zaibatsu*, there was stiff competition

among a large number of financial institutions. In the late 1920's, the Ministry of Finance successfully promoted regulations designed to foster concentration in banking, but it was only after wartime policies explicitly linked firms with banks that the now-familiar main-bank system emerged. Under the financial system of the 1950's, 1960's, and 1970's, a relatively small number of banks used the market power conferred on them by the Ministry of Finance to transfer the savings of their household depositors at less than competitive rates of interest to Japan's rapidly growing finance-short industrial sector. Bank finance replaced the direct financing that had predominated earlier in the 20th century.

The increasing regulation of the financial system in the mid-1920's went hand-in-hand with increasingly active involvement of the government in shaping Japan's industrial structure. However significant the role of the Japanese government may have been in any particular industry, either in the 1920's or before, only a decade later, under the pressure of Japan's increasing military involvement on the China mainland, did comprehensive national economic planning emerge in Japan. In the 1950's, 1960's, and 1970's, the Japanese government's role in shaping the economy's structure came to resemble neither its wartime role nor the more limited role it had played earlier in the 20th century. Comprehensive national economic planning remained as a wartime legacy, but the Japanese government's role was indicative and market-driven. The major policy instruments the government retained to implement sectoral policy included (i) the preferential allocation of foreign exchange and import licenses, (ii) discriminatory tax-subsidy provisions and import tariffs, and (iii) subsidized loans from government financial institutions and implicit influence on the allocation of loans made by the heavily regulated private financial sector.

III. Japan's Distinctive Postwar Economic System

The economic institutions that characterized Japan in the 1950's, 1960's, and 1970's are generally regarded as less similar to Western economic institutions than the economic

institutions that characterized Japan earlier in the 20th century. The economic institutions of Japan's postwar system can be thought of as mutually reinforcing. The concentration of financial power in Japan's banking system, so long the goal of the Ministry of Finance, greatly facilitated (but also made necessary) an active, indicative, sectoral policy on the part of the Japanese government. Without countervailing government pressure, in the absence of market discipline, the complicated pressures of intra-bank group politics might well have skewed resource allocation in the direction of established mature industries. At the same time, absent a group of powerful banks who maintained long-term relationships with the firms who borrowed from them, it is unlikely that many Japanese firms would have been willing to risk the financial inflexibility associated with a permanent employment system.

Much has been claimed on behalf of the Japanese economic system as it operated during the 1950's, 1960's, and 1970's. Bank-oriented financial control and permanent employment, among other Japanese economic institutions, are said to be unusually effective in overcoming the problems of adverse selection and moral hazard endemic in virtually all economic systems. Likewise, the Japanese government's role in the economy is said to have raised the share of resources being devoted to capital formation even as it played an implicit coordinating role in overcoming the market failures that inhibit structural transformation.

To date, such claims about postwar Japan's economic system have fared better at the theoretical level than at the empirical level. Quite apart from the absence of any clear indication that the aggregate performance of the Japanese economy was relatively stronger, by comparison with other advanced industrial economies, in the 1950's, 1960's, and 1970's than in earlier or in subsequent periods, the systematic statistical investigations that have been done at less aggregated levels yield mixed evidence at best. For example, there are careful statistical studies that suggest that the net impact of the Japanese government's industrial policies may have been no more than the support of declining industries without much impact on aggregate productivity (Richard

Beason and David Weinstein, 1996). Likewise, there are several careful statistical studies that find main banks altering the behavior of their affiliated firms without necessarily improving their performance (see e.g., Akiyoshi, Horiuchi et al., 1988).

IV. Systemic Evolution and Japan's Economic Distress

Inevitably, much of the analysis of the distinctive economic institutions characterizing Japan's era of high growth between 1953 and 1973 occurred after 1973, when these institutions were already undergoing significant transformation. High leverage with diversification and bank-monitoring may well be a good recipe to force corporate managers to perform efficiently, but it may have relatively little to do with the Japan of the 1980's and 1990's. The deregulation of Japanese financial markets which began in the late 1970's, and particularly the removal of controls on the inflow and outflow of capital in the early 1980's, long ago changed the structure and operation of the Japanese economy. Slower growth and the freedom to raise funds in international (and later in domestic) markets meant a significant loosening of the relationship between many of Japan's best-known and best-run corporations and their banks and a correspondingly greater role for equity financing for such firms.

The same financial deregulation that allowed Japanese firms to draw on far more diverse sources of finance changed the role the Japanese government played in the economy. In the 1980's, Japanese firms seeking to promote new industries rarely needed the Japanese government as an ally to force a bank to turn on its financial spigot. The same deregulation that removed the need for the government to intervene also removed the means by which the government might intervene. The Japanese banking system, now forced to compete with many other sources of finance both at home and abroad, long ago ceased to be an effective instrument with which to shape Japan's industrial structure.

While Japanese economic institutions have changed in significant ways since the early postwar decades, the prolonged downturn of the 1990's has engendered the widespread belief that

further adaptation is necessary if Japan is to improve its economic performance. Most attention continues to be focused on Japan's financial system. Twenty years of financial deregulation have made it significantly easier for the traditional borrowers to find new sources of finance without making it correspondingly easier for banks to maintain the quality of their loan portfolio. By way of compensation, until recently, banks were protected by the Ministry of Finance from the consequences of this asymmetric deregulation. This new moral hazard left Japanese banks in the 1990's with an extraordinary number of nonperforming loans on their balance sheets. To remove the source of the banking sectors' problems, in November 1996 Prime Minister Ryūtarō Hashimoto proposed a five-year deregulation plan designed to make it easier for banks to enter markets traditionally reserved for nonbank financial institutions, even as the way is eased for these institutions to enter markets traditionally reserved for banks. This five-year financial deregulation package, which is part of a much larger deregulation program, is designed to make Japanese financial services more competitive. The intended increased securitization of Japanese finance and new information-disclosure requirements should increase the transparency of the Japanese firm. With better information, Japanese financial markets should be better able to concentrate large, much-needed resources in new areas of development.

While deregulation is changing the character of Japanese capital and product markets, demography is altering Japanese labor markets even as it creates enormous uncertainty about the future of Japanese saving relationships. Increasing labor scarcity brought on by the aging of the Japanese population will allow employees more say in their terms of employment (e.g., flexibility in working hours and age of retirement) than has been true in the past. Whether permanent employment remains, even as Japan's labor force begins an historic decline in its absolute size, is quite another matter.

While the productivity of the Japanese economy remains inferior to the global standard in many areas, by comparison with even the recent past, Japan is close enough to the technological frontier that determining the technological trajectory that a firm will follow is far more complicated than it once was. In this environment, with

future human-resource needs difficult to predict, Japanese firms of the future may well prefer to have their labor force bear more both of the risks associated with acquiring specialized training and the risks associated with secular and cyclical demand shocks. Such steps will require a change in the way in which training is provided and changes in the Japanese government's education and social policies. New graduate schools and vocational schools, together with a different menu of social-security programs, will be needed.

It is likely Japan will enter the 21st century with economic institutions increasingly unlike those institutions associated with Japan's superior economic performance in the last half of the 20th century. If the past hundred years provide cause for optimism, new economic arrangements, unhappily now only dimly visible, among firms, labor, finance, and government and between the Japanese economy and the global economy will work to mitigate the consequences of the changed demographic and technological environment facing the Japanese economy. Following improvements in Japan's growth performance relative to the industrialized economies of Western Europe and North America early in the 20th century, Japan's rate of economic growth has been dominated by external shocks common to all these economies. It is not obvious that the 21st century will be different.

REFERENCES

- Beason, Richard and Weinstein, David. "Growth Economies of Scale and Targeting in Japan (1955-1990)." *Review of Economics and Statistics*, May 1996, 78(2), pp. 286-95.
- Horiuchi, Akiyoshi; Packer, Frank and Fukuda, Shin'ichi. "What Role Has the 'Main Bank' Played in Japan?" *Journal of the Japanese and International Economics*, June 1988, 2(2), pp. 159-90.
- Maddison, Angus. *Monitoring the world economy*. Paris: Organization for Economic Cooperation and Development, Development Centre, 1995.
- Mankiw, N. Gregory. "The Growth of Nations." *Brookings Papers in Economic Activity*, 1995, (1), pp. 275-310.
- Saxonhouse, Gary R. "Productivity Change and Labor Absorption in Japanese Cotton Spinning, 1891-1935." *Quarterly Journal of Economics*, May 1977, 91(2), pp. 195-219.

Declining Population and Sustained Economic Growth: Can They Coexist?

By YUTAKA KOSAI, JUN SAITO, AND NAOHIRO YASHIRO*

Recently published official population projections suggest a steady decline in the Japanese population after the first decade of the 21st century, mainly due to a continuous decline in the fertility rate. Though a declining population and labor force would suggest lower economic growth, the constraint by itself would stimulate efficiency-augmenting technological changes by tightening labor and capital resource constraints. Also, as a result of changes and reforms in the labor market, socioeconomic forces can work to mitigate the decline in fertility. This paper discusses Japan's growth prospects in the 21st century, focusing on the endogenous mechanism countering the declining population, and draws lessons from the historical experiences of Japan as well as other countries.

I. Past Economic Development in Japan and East Asia

The celebrated "East Asian Miracle" (World Bank, 1993) was intellectually challenged by Paul Krugman (1994). According to his view, the high economic growth in the region was not the result of "gains in efficiency" or total factor productivity (TFP) growth as in many of the OECD countries, but was instead mainly the result of an "input-driven growth" just like that in the former Soviet Union in the 1960's. Thus, the Asian growth will eventually cease, with its labor endowments being fully employed and its mobilization within the economy coming to an end.

However, this view neglects the possibility that a country might shift from input-driven growth to that associated with gains in efficiency in accordance with changes in the economic environment. It is advantageous for Asian countries to pursue resource-utilizing economic growth as long as they are abundantly endowed with factors of production in the country. Even the United States, according to one estimate, took the input-driven growth pattern at its initial stage of economic development: U.S. GDP growth in the 1820–1870 period was 4.22 percent while the contribution of TFP growth was –0.15 percent (Angus Maddison, 1995). The failure of the former Soviet Union in shifting from input-driven growth to that of gains in efficiency was mainly due to its insufficient structural adjustments to changing factor endowments, owing to the centrally planned economic system. The government's interventions in most East Asian economies are "market-friendly" ones in contrast to the market-repressing policies in the former Soviet Union (Takatoshi Itoh, 1996).

Japan's economic development provides a clue to the transformation of the growth pattern. We divide Japan's economic growth in the last hundred years into four periods. The average rate of economic growth in the prewar period was about 3 percent, and the TFP growth in the period after the World War I was slightly higher than in the preceding period. The most significant period came right after World War II. The markedly high economic growth, on average 8 percent for two decades up to the mid-1970's, was associated with drastic changes in the economy and society preceding the East Asian Miracle (Kosai, 1986). The high-growth era ended with the catching-up of Japan's industry to the level of other OECD countries in the mid-1970's. According to the analysis based on the growth-accounting method, the contribution rate of TFP to total GDP growth has steadily risen over the four periods from 20 percent to 47

* Kosai and Saito: Japan Center for Economic Research, 2-6-1 Nihonbashi-Kayabacho, Chuoku, Tokyo, 103-0025, Japan; Yashiro, Institute of International Relations Sophia University, 7-1 Kioicho, Chiyodaku, Tokyo 102-0094, Japan. We thank Hugh Patrick and David Weinstein for their helpful comments and JCER for financial support. Supporting figures and tables are available from authors on request.

percent. Thus, Japan's case clearly provides counterevidence to Krugman's hypothesis that resource constraint would put an end to input-driven growth.

II. Causes and Consequences of the Declining Population

A major resource constraint in the coming decades is the declining population. The total fertility rate (average number of children per woman, over her lifetime) in Japan fell from the postwar peak of 4.5 in 1947 to 2.1 in the 1950's, stabilized there for two decades, and then fell again from the mid-1970's to 1.4 in 1996, which is far below the population reproduction rate. The decline in fertility is mainly attributable to an increase in the opportunity cost of women's child-rearing. The ratio of self-employed women in total employment fell, while that of paid-employment rose from 15 percent to close to 40 percent between 1955 and 1996, reflecting rapid economic growth. The improved job opportunities for women increased the "trade-off" facing them (i.e., that between job continuation and child-bearing). It is particularly so under the fixed employment practices in Japan in which the reentry of women after child-rearing to good job markets is highly restricted. This has led to lower marriage rates, falling from 82 percent to 52 percent at age 25–29 between 1970 and 1995.

Even with an optimistic assumption that the fertility rate will eventually recover after 2000 to stabilize at 1.6, the population of those between ages 15 and 64 will decline by 17 percent from the 1995 level by 2025. Also, Japan has the highest longevity in the world, reflecting the high level of per capita income and health-care services. As a result, the ratio of the Japanese elderly in the total Japanese population will be 27.4 percent in 2025, which would be among the highest in OECD countries.

In projecting the economic growth in the first half of the 21st century, a key issue is how to account for the endogenous mechanisms offsetting the negative impacts from the declining population. Several previous models have derived a negative relationship between fertility and the narrowing male–female wage gap (Oded Galor and David N. Weil, 1996).

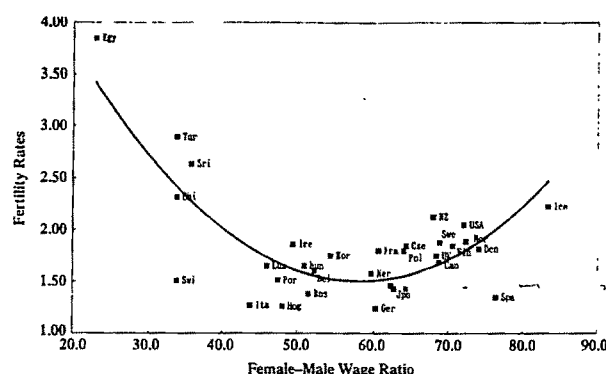


FIGURE 1. FEMALE–MALE WAGE RATIO AND FERTILITY RATE

Nevertheless, a cross-country comparison of major OECD countries shows that those countries where women's wages are closer to men's generally have higher fertility. This implies that fertility rates decline initially with higher wages of women relative to men's, and the associated higher opportunity costs for child-bearing (substitution effect). But with an increasing scarcity of the labor supply as the result of the declining population, job opportunities and wages between the genders become more equal, and the income effect with respect to the demand for children comes to dominate. Thus, the catching-up of women's wages to that of men's can coexist with the recovery of the fertility beyond a certain point. This is consistent with the observed "U-shaped" association between the ratio of women's wages relative to men's and fertility rates when we pool both OECD and developing countries together (Fig. 1).¹

The U-shaped association between the wage disparity and fertility across countries suggests that the differences in factor endowments between the sexes are not innate, but more a matter of social practices, and they become less important with economic and social development as a result of the following three factors:

¹ Similar relationships in terms of male–female labor-force participation cannot be found, mainly due to a high incidence of self-employed in developing countries. Among OECD countries, there is a statistically significant result that a 10-percent difference in women's labor-force participation corresponds to a change of 0.12 in the fertility rate.

(i) an increase in investment in women's higher education; (ii) the lower cost of women's child-rearing and other home production by their being shared more equally between men and women; (iii) the more equal employment opportunities due to the flattening wage curve by age and longer maternity leaves, as well as better child-care services.²

Also, economic growth could be sustained even under a declining population if it is supported by an endogenously induced technological progress in the market (i.e., if the tightening labor-market conditions tend to stimulate better utilization of the scarce resources). Similar tendencies are observed in the OECD countries; those countries with less-abundant labor forces tend to have higher labor-productivity growth (David M. Cutler et al., 1990), which could mostly be attributed to TFP growth (Yashiro and Akiko Oishi, 1997). The mechanism underlying this endogenous process of TFP growth is interpreted as follows. Accumulation of human capital and a decline in the labor force would raise real wages at a faster pace than in other countries, reflecting the different rates at which the respective societies are aging. It would also increase the return from utilizing human capital and would thereby have the effect of stimulating the development of innovation, which would lead to human-capital-intensive technologies (Yujiro Hayami, 1997). These outcomes would also be facilitated by positive external factors, namely, the increasing pool of human capital and the consequent decline in the cost of technological development.

We present a macroeconomic model which incorporates these endogenous mechanisms of the declining population. The estimation is based on the data between 1947 and 1996, and the simultaneously solved re-

sults are used to project up to the year 2025.³ In our macroeconomic model, the fertility rate, which determines the population growth, is set as a function of child mortality, the ratio of women to men's employment (excluding the self-employed), and the availability of child-care services. Both women and men's employment as ratios to their respective populations are functions of labor productivity. Labor productivity (as well as GDP) is derived from the capital-labor ratio adjusted by the capacity-utilization ratio (exogenous variable); the latter is also used as indicating the efficiency of the market in the simulations. The labor force is derived from population statistics, and capital stocks are derived from the rate of return on capital, which is approximated by the ratio of GDP to capital stocks; the national saving ratio is used as an indicator of financial costs. The national saving ratio is formulated based on the ratio of total employment to total population. Population is derived from fertility and mortality rates (exogenous factor).

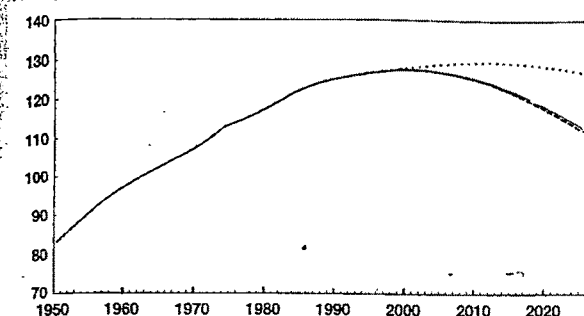
We compare the baseline case with the cases when endogenously stimulated counteracting forces are accounted for. In the baseline case, the fertility rate will continue to decline until it stabilizes at quite a low level. An increase in the share of the retired in the total population would be associated with falling ratios of household savings. Also, a declining labor force by itself lowers capital profitability. Thus, both declining capital profitability and higher costs for capital would decrease investment as a proportion of GDP. Per capita GDP will fall slightly beyond the year 2010, mainly because the declining labor force and saving ratio would discourage investment and lower the capital-output ratio. Real GDP in 2025 will be lower than that in 2010 by 10 percent (an average growth of -0.7 percent during the period).

In comparative case 1, we assume that the tightening labor-market conditions, being supported by deregulation and other competition-

² In Japan, the wage gap between the sexes is mainly attributable to the sharp age-wage profile of men and insufficient mid-career recruitment opportunities of women. With the flattening of the wage profile as a result of the aging of the work force, the employment opportunities for women after child-rearing ages could be expanded, thereby lowering the opportunity costs of leaving the labor market due to child-bearing.

³ This model is of a closed-economy type. The expansion of the model into an open-economy type with multiple goods is likely to mitigate the negative impacts from declining population.

A. Population (million)



B. GDP (trillion yen, constant prices)

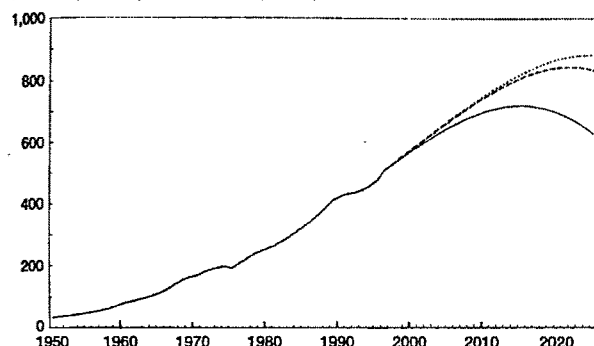


FIGURE 2. MAJOR SIMULATION RESULTS: (A) POPULATION (MILLIONS); (B) GDP (TRILLIONS OF YEN, CONSTANT PRICES)

Key: Solid line, baseline; dashed line, simulation case 1; dotted line, simulation case 2.

stimulating measures, would lead to endogenously stimulated gains in efficiency at a level equivalent to the average of the late 1980's. As a result, labor productivity would be higher, and the saving ratio would rise with the increase in employment. This would bring about a slight (0.7 percent) increase in real GDP between 2010 and 2025.

In comparative case 2, we add another assumption to that in the case 1: the negative effect of increasing employment of women relative to men on fertility will be less, mainly due to more participation of men in home production and to labor-market reforms aimed at lowering the barriers for working women with small children to reenter the labor market. These effects are approximated by gradually reducing the parameter of the gender employment gap in the equation explaining fertility rate by one-fourth between the present and 2025. This would help the fertility rate to re-

cover to the level of 2.0, which is close to the level stabilizing the population. Women's employment will be stimulated, with the total employment basically maintained at the same level up to 2025. As a result, the national saving ratio would fall less dramatically than in the baseline, and real GDP growth between 2010 and 2025 would be an annual average of 1.0 percent (Fig. 2).

III. Conclusion

The negative consequences of population decline can be avoided, if endogenously induced mechanisms are sufficiently effective. An increasing scarcity of labor would stimulate more efficient utilization of resources, shifting the economic growth pattern from the "input-drive type" to that of "gains in efficiency." The fertility rate could also recover if it is supported by labor-market reforms that ease the trade-off between women's staying on the job and their child-rearing. Thus, projecting Japan's long-term economic growth is conditional to the extent that these counteracting forces in the market can be enhanced by policies for liberalizing various regulations in the labor market.

REFERENCES

- Cutler, David M.; Poterba, James M.; Sheiner, Louise M. and Summers, Lawrence H. "An Aging Society: Opportunity or Challenge?" *Brookings Paper on Economic Activity*, 1990, (1), pp. 1-73.
- Galor, Oded and Weil, David N. "The Gender Gap, Fertility, and Growth." *American Economic Review*, June 1996, 86(3), pp. 374-87.
- Hayami, Yujiro. *Development economics*. Oxford: Oxford University Press, 1997.
- Itoh, Takatoshi. "Japan and the Asian Economies: A Miracle in Transition." *Brookings Paper on Economic Activity*, 1996, (2), pp. 205-72.
- Kosai, Yutaka. *The era of high-speed growth*. Tokyo: University of Tokyo Press, 1986.
- Krugman, Paul. "The Myth of Asia's Miracle." *Foreign Affairs*, September/October 1994, 73(6), pp. 62-78.

Maddison, Angus. *Monitoring the world economy, 1820–1992*. Paris: Organization for Economic Cooperation and Development, 1995.

World Bank. *The East Asian miracle*. Oxford: Oxford University Press, 1993.

Yashiro, Naohiro and Oishi, Akiko. "Population Aging and the Saving–Investment Balance in Japan," in Michael Hurd and Naohiro Yashiro, eds., *The economic effects of aging in the United States and Japan*. Chicago: University of Chicago Press, 1997, pp. 59–88.

The Incentive Structure of a "Managed Market Economy": Can It Survive the Millennium?

By KOICHI HAMADA *

The 20th century was a tempestuous century for Japan. Imagine a period of time, back about a hundred years ago, when Japan had just taken off on its path of modern economic growth. At that time, most of the signs that would foretell the future of the Japanese economy were already operative. In 1897, Japan's cotton export exceeded its cotton import for the first time in history. In 1900, environmental hazards from water pollution caused by the Ashio copper mine triggered a protest that was led by Diet Member, Shozo Tanaka. In that same year, the government legislated the Policing of the Public Order Act (*Chian Keisatsu Ho*), a precursor of the notorious Maintenance of the Public Order Act (*Chian Iji Ho*), which suppressed democracy in Japan. These and other characteristics were indicators of what was to come in Japan's economic future.

In the first half of the 20th century, as a latecomer to world markets, Japan imitated the imperialistic expansions of Western countries. The ensuing reckless attempt to reign over the Pacific resulted in complete failure, when Japan lost World War II in 1945.

In the second half of the 20th century, Japan completely shifted gears and veered toward pursuing economic prosperity as a merchant-marine country. Through the "era of high-speed growth (Yutaka Kosai, 1986)," Japan realized the "miracle of the Rising Sun" to become "Asia's new giant" (Hugh T. Patrick and Henry Rosovsky, 1976). Even after the reflection point caused by the first oil crisis during 1973-1974, Japanese industries provided high-quality goods for world markets. Japan accumulated a sub-

stantial balance-of-payments surplus and emerged as a major supplier of capital in the world. This short paper focuses on the incentive mechanisms that originally led to the success of the Japanese economy, but which now have created problems.

I. A Half-Century as a Merchant-Marine Country

The remarkable growth trajectory of the postwar growth was interrupted by the oil crisis during 1973-1974, a crisis that became an inflection point in Japan's rapid growth. In spite of that downturn, Japanese manufactured goods were welcomed by world markets, and helped by the high savings rate, Japan grew to be a large capital-export country. At the end of 1992, Japan possessed \$513 billion worth of external assets. In spite of the major quantitative role of Japan as a world creditor country, however, one could not help but wonder if the quality of her financial-service industries was really first class (Hamada, 1994).

From the beginning of this decade, the Tokyo Stock Exchange started to decline. Delinquent debts of banks and other private as well as public financial institutions began to become heavy burdens on the Japanese economy. Quite recently, the Yamaichi Securities Co. and the Hokkaido Takushoku Bank were declared bankrupt. The last decade of Japan's century, so to speak, was thus tainted by financial turmoil.

Behind this sequence of financial collapses, the incentive problems existed, as will be explained below. At the same time, one should not neglect the fact that monetary policy amplified, rather than smoothed, business fluctuations. Japan went into a severe recession after a series of policy mistakes: excess liquidity (1988-1989), a sudden contraction of the money supply (1990), and tax raises

* Economic Growth Center, Yale University, 21 Hillhouse Avenue, New Haven, CT 06511. I am indebted to David Weinstein and Fumiko Takeda for their insightful comments and to Carolyn M. Beaudin for her editorial contributions.

that discouraged consumer spending (1997) (see Hamada, 1995).

II. The Incentive Structure That Sustained as Well as Jeopardized the Japanese System

Until recently, the incentive structure of the Japanese economy was something to analyze, commend, and imitate from abroad. Japanese firms, administrative guidance, homogeneous education, family structure of the society, Total Quality Circles (TQC's), the Just-in-Time (*Kanban*) System, and the main bank system (Masahiko Aoki and Patrick, 1994) all were considered to have helped the Japanese miracle during its high-growth period and beyond. In fact, the incentive structure was studied at the world level as a model for developing countries (World Bank, 1993).

As a result of the financial turmoil during the 1990's, the Japanese incentive structure is now under scrutiny. Consider the incentive constraints that an economic agent in Japan faced. In most of the goods and services markets, the price mechanism was the basic guiding force for resource allocation. At the same time, community pressures to conform to a group norm were coming from within the firm, from business liaisons (*Keiretsu*), and from the government bureau in charge of the industry (*Genkyoyu*). Administrative guidance that hardly had any explicit legal basis intervened in all sectors of the economy.

During the second half of the 20th century, the Japanese economy was influenced by a combination of the market mechanism and the formal and informal guidance of the government. One could call it a "managed market economy." An interesting hypothesis traces the origin of this type of economy to the preparation process for and practices used during World War II, that is, the "1940's theory" of the origin of Japanese characteristics (Tetsuji Okazaki and Masahiro Okuno, 1993; Yukio Noguchi, 1995).

In order to mobilize materials, goods, services, and arms for the war objectives, administrative controls and guidance were apparently more effective than were price signals representing a relative scarcity of goods and services. According to this 1940's theory,

most of the main features that distinguish the Japanese economy were embedded in the process of preparing for participating in World War II. Those features are often attributed to the indigenous community structure of the traditional society, but the theory argues that most of them are the products of the institutional adjustment to mobilize the economy for procurement for the war.

Many puzzles remain. Even though they might be products of a wartime economy, why did these systems survive without being absorbed into an individualistic society like those of Western countries? Were the community consciousness and practices effective means for sustaining those systems long after the war?

Under this managed market economy, economic agents had to keep their eyes not only on what was occurring in the markets, but also on what would be the next move of supervising ministries. Without government-oriented mobilization, preparing and executing military activities would have been difficult. In the postwar period in Japan, when such needs disappeared, the system, with close interventions from the government, continued nevertheless.

In industrial sectors, interventions from government gradually subsided. After the liberalization of automobile imports which took place around the middle of the 1960's, few protective measures remained in industrial sectors. On the other hand, in service sectors, particularly in the financial sectors, the situation is the opposite. Until the 1980's, deposit interest rates were strictly regulated by a kind of "Regulation Q," and financial institutions were not allowed to develop new financial instruments without the approval of the Ministry of Finance (MOF). Bankers did not heed the signals of markets as much as they did signals for the next policy moves of administrative agencies. In any bank, a capable middle manager was assigned as an "MOF *tan*" and a "Bank of Japan *tan*," where *tan* means a contact or a watch person. The assigned watch person to the MOF frequented the Banking Department of the Ministry, socialized with officials, and made the best efforts to obtain information about the future course of the MOF policy with regard to the banking sector.

(Okuno, 1993). Financial efficiency was hard to attain, and it was possible for a hotbed for corruption to simmer.

As a consequence, one now sees pervasive difficulties in the financial-service sectors. Many of the residential financial institutions (*Jusen*) and credit unions have gone bankrupt. The Hokkaido Takushoku Bank is out of business, and so is Yamaichi Securities, the fourth largest brokerage house in Japan.

It is interesting to contrast the difference in actions taken in the two crises that hit Yamaichi approximately 30 years apart. In the first crisis, which was triggered by a newspaper leak in 1965,¹ the Bank of Japan engaged in a rescue operation by the emergency lending of ¥28.2 billion (\$78 million). In the recent crisis in 1997, the government decided to let the company go bankrupt after the hidden loss of ¥260 billion (\$2 billion) was revealed.

In 1965, Japan was in a short recession in the middle of her rapid economic growth. Private bankers were persuaded by the government to contribute to the rescue fund. The Japanese financial market was not under fierce international competition, and the government had the leadership and accountability to rescue Yamaichi. In fact, the Ministry of Finance gained more power by establishing the system that requires *approval* rather than registration for opening a security company.

Thirty-two years later, Japan was in a long recession that had lasted seven years already. The Fuji Bank, the main bank of Yamaichi, is reported to have refused to help Yamaichi because of its own difficulties. The Japanese capital market is liberalized so much that the Japanese monetary policy cannot alone influence financial variables. The MOF has lost the power, as well as its credibility, for reorganizing Japan's financial system in the future.

III. Are Asian Miracles or Mirages Evaporating?

During the last quarter of this century, economies in East and Southeast Asia achieved re-

markable growth. Since government controls are imposed upon market mechanisms in many of those countries, the growth phenomenon was interpreted to be a transplant of the Japanese system and was referred to as the "East Asian Miracle" (World Bank, 1993). Paul Krugman (1994) downplayed this phenomenon on the grounds that the growth of Asian countries, if not of Japan, was a process of increasing inputs rather than increasing productivity. (For a more optimistic view, see Steven Radelet and Jeffrey Sachs [1997].) In my opinion, the process of increasing employment of productive factors itself should be highly appreciated. Low-income countries all over the world suffering from unemployment problems would greatly appreciate being able to achieve growth through more employment of resources!

Unfortunately, the incentive mechanisms in financial markets in these countries are in just as much trouble, or perhaps even more, as those in Japan in the 1990's. Does this mean that the Japanese incentive mechanism has revealed its weakness in other Asian countries as well? Tentative answers are as follows.

First, though the mixture of government controls and the market mechanism is found in all these countries, their development patterns are diverse (Takatoshi Ito, 1996). There is no assurance that the same incentive mechanism was working in these countries. Second, financial difficulties are caused by macroeconomic policy mistakes as well as microeconomic incentive problems. Asian countries may have been the victims of improper macroeconomic policies, as was Japan. No incentive mechanism is free of fault when severe macroeconomic turbulence exceeds the tolerance limit of the system. Third, these countries have problems because strong government controls and open capital markets do not get along, and under an open capital market, an independent monetary policy and a fixed exchange rate cannot coexist.

The proposal of Japan to establish an "Asian Monetary Fund" to rescue Asian economies is an indication that Japan is able to assume a leadership role in the world

¹ The MOF almost controlled the news media as well, but the control was not perfect.

economy. Interestingly, the proposal triggered negative reactions from the United States, Europe, and the IMF. Since Japan complains of its lack of influence relative to the magnitude of its financial contributions to the world (and to bailout funds for ailing Asian economies), it is natural for Japan to seek to increase its leadership by establishing such a fund. International organizations like the IMF are opposed to the establishment of any independent, money-lending, regional institution. Also, critics from the United States and European countries stress the concern that an international organization operated by Japan-like principles may become too lenient to problem countries, and that the moral hazard of domestic as well as international lendings may develop.

IV. Concluding Observations

Aside from long-term problems like aging, Japan faces deep troubles in the financial sectors. Failure in incentive mechanisms is apparent. Japan is not just at the turn of the century—the Japanese incentive system is currently at a turning point. Will the century of the Rising Sun be over in a few years?

I will finish the paper with some optimistic notes. When the U.S. automobile industry had difficulty because of Japanese competition, the United States enticed Japanese foreign investments and subsequently learned from Japan about improving productivity and quality of products to “come back” (Paul Ingrassia and Joseph B. White, 1994). In the 21st century, I hope, Japan will learn from other countries. With an open economy, after the “Big Bang,” the Japanese, who are not intrinsically weak in finance, might invite participation from foreign financial institutions and learn modern financial entrepreneurship from the nearby “display windows.” The recent appeal to the World Trade Organization regarding the Kodak–Fuji case illustrates the new Japanese attitudes of articulation and legal orientation that are necessary for successful financial business. Every cloud has a silver lining.

REFERENCES

- Aoki, Masahiko and Patrick, Hugh T., eds. *The Japanese main bank system: Its relevance for developing and transforming economies*. Oxford: Oxford University Press, 1994.
- Hamada, Koichi. “Japan’s Prospective Role in the International Monetary Regime,” in C. C. Garby and M. B. Bullock, eds., *Japan: A new kind of super power*. Baltimore, MD: Johns Hopkins University Press, 1994, pp. 143–58.
- . “Bubbles, Bursts, and Bailouts: Comparison of Three Episodes of Financial Crises in Japan,” in M. Okabe, ed., *The structure of the Japanese economy: Changes in the domestic and international character*. New York: St. Martin’s, 1995, pp. 263–86.
- Ingrassia, Paul and White, Joseph B. *Comeback: The fall and rise of the American automobile industry*. New York: Simon and Schuster, 1994.
- Ito, Takatoshi. “Japan and Asian Economies: ‘A Miracle’ in Transition,” *Brookings Papers on Economic Activity*, Fall 1996, (2), pp. 205–73.
- Kosai, Yutaka. *The era of high-speed growth: Notes on the postwar Japanese economy*. [Jacqueline Kaminski, translator]. Tokyo: University of Tokyo Press, 1986.
- Krugman, Paul. “The Myth of Asia’s Miracle,” *Foreign Affairs*, November/December 1994, 73(6), pp. 62–79.
- Noguchi, Yukio. *1940 Nen-taisei: Saraba “senji keizai”* [The 1940 regime: Good bye to the “war-time economy”]. Tokyo: Toyo Keizai Shinposha, 1995 (in Japanese).
- Okazaki, Tetsuji and Okuno, Masahiro. “Gendai Nihon no Keizai Shisutemu to sono Rekishiteki Genryu [The Modern Japanese Economic System and its Historical Origin],” in Tetsuji Okazaki and Masahiro Okuno, eds., *The origin of the modern Japanese economic system*. Tokyo: Nihonkeizai, 1993, pp. 1–34 (in Japanese).
- Okuno, Masahiro. “Gendai Nihon no Keizai Shisutemui Sono Kozoto Henkaku no Kanosei

- [The Modern Japanese Economic System: Its Incentive Structure],” in Tetsuji Okazaki and Masahiro Okuno, eds., *The origin of the modern Japanese economic system*. Tokyo: Nihon-keizai, 1993, pp. 273–91 (in Japanese).
- Patrick, Hugh T. and Rosovsky, Henry. *Asia's new giant*. Washington, DC: Brookings Institution, 1976.
- Radelet, Steven and Sachs, Jeffrey. “Asia's Re-emergence.” *Foreign Affairs*, November/December 1997, 76(6), pp. 44–59.
- World Bank. *The East Asian miracle: Economic growth and public policy*. New York: Oxford University Press, 1993.

Competition, Policy Burdens, and State-Owned Enterprise Reform

By JUSTIN YIFU LIN, FANG CAI, AND ZHOU LI*

One of the most important remaining issues in China's transition to a market economy is the reform of state-owned enterprises (SOE's). When reforms started in late 1978, SOE's dominated China's industrial sectors in every aspect. After 18 years of gradual transition, the SOE share in China's total industrial output has declined from 77.6 percent in 1978 to 28.8 percent in 1996. However, in 1996 SOE's still employed 57.4 percent of urban workers and possessed 52.2 percent of total investment in industrial fixed assets. Improving SOE performance is crucial for social stability and sustained growth in China. However, over 40 percent of SOE's are losing money. In this paper, we will argue that the root of the SOE problem is the separation of ownership and control and that the often-criticized soft-budget constraints arise from various state-imposed policy burdens, which make the state accountable for the poor performance of SOE's. The key for a successful SOE reform is to remove the policy burdens and to create a level playing field so that market competition can provide sufficient information for the managerial performance of the SOE's and make the managers' incentives compatible with those of the state.¹

[†] *Discussants:* Barry Naughton, University of California—San Diego; Loren L. Brandt, Jr., University of Toronto.

* Lin: China Center for Economic Research, Peking University, Beijing, 100871, China, and Department of Economics, Hong Kong University of Science and Technology, Hong Kong; Fang Cai and Zhou Li: Chinese Academy of Social Sciences, Beijing 100732, China.

¹ In this paper we limit our discussions to the reform of large-scale SOE's, which consisted of 5.6 percent of SOE's in terms of number of enterprises in 1996 but contributed to 63.3 percent of the gross output value of SOE's in the same year. For the medium- and small-scale SOE's, lease, privatization, or bankruptcy are appropriate reform programs.

I. Competition and the Performance of Large Corporation in a Market Economy

Although, by definition, the state owns the SOE's, the state cannot operate them by itself and needs to delegate their control to the enterprises' managers. The separation of ownership and control is a common feature of any large modern corporation. Due to this separation, the issues of incentive incompatibility and information asymmetry often arise between the managers and the owners. Agency problems, such as the moral hazard and managerial slacks and discretion, may surface. The success of any large modern corporate institution depends on its ability to overcome these problems. Intuitively, one possible way out for the owners is to oversee the managers' actions directly and to reward the managers according to their managerial efforts (Armen A. Alchian and Harrod Demsetz, 1972). In practice, total observation of managerial actions in a large modern corporation is either impossible or prohibitively costly. Moreover, the owners of a large modern corporation are numerous. Due to the free-rider problem, any individual owner of a firm will not have the incentive to oversee the detailed activities of the firm. The large modern corporation does not have owners in the same sense as in the property-rights literature.²

The prevalence of large modern corporations in the market economy indicates that some institutional arrangements to mitigate the agency problems must exist. The recent liter-

² An institutional investor may own a large share of a firm. However, the institutional investor is also an agent who may not have the right incentives to monitor the managers. The owners of the institution's funds may not have incentives to monitor the institutional investor either.

ature suggests that market competition is such an arrangement. Some summary indicators in a competitive market, such as relative profits of firms in the market, provide a sufficient-statistic condition for evaluating managers' performance (Bengt Holmstrom, 1982). With the sufficient statistic, incentive compatibility between the owners and the managers can be achieved in two ways: directly, the owners can design a managerial-compensation scheme that is based on the comparison of the firm's performance with the industrial average or on the rank of the firm's performance in the industry (Holmstrom, 1982); indirectly, a firm's performance in a competitive market provides a signal to the managerial labor market about the manager's talent and behavior, and the signal determines the manager's future wages (Eugene F. Fama, 1980).³

II. Endogeneity of the SOE Management Institution in a Soviet-Type Economy

A salient feature of traditional SOE's in the Chinese economy and other Soviet-type economies was their lack of autonomy. The state provided all inputs to SOE's for their production according to central plans and covered all their costs. In turn, the SOE's delivered to the state all outputs and revenues. The state set the wage rates of the SOE workers and managers. All activities of the SOE's required the state's approvals. Such a seemingly irrational arrangement in effect was an endogenous response to the agency problems in a traditional Soviet-type economy.

The Soviet-type economy is known for its maximal mobilization of resources for the establishment of capital-intensive heavy industries. These heavy-industry projects had three characteristics: (i) they required a long gestation; (ii) for a less-developed country, most equipment for the projects needed to be im-

ported from developed countries; and (iii) each project required a large lump-sum investment. However, the Soviet-type economies were built in low-income agrarian countries which also had three characteristics: (i) capital was scarce and the market-determined interest was high; (ii) exportable goods were limited, and the market-determined price for foreign exchange was expensive; and (iii) funds for large projects were hard to mobilize because economic surplus was small and scattered. For the purpose of reducing the costs and mobilizing funds for the heavy-industry projects, the Soviet-type economy formed a distorted macroeconomic-policy environment in which interest rates, foreign-exchange rates, wage rates, and prices for raw materials and other products were artificially suppressed (Lin et al., 1996).

The above macroeconomic-policy distortions induced a total imbalance in the supply and demand for credits, foreign exchanges, raw materials, and other products. Because nonpriority sectors were competing with priority sectors for the low-priced resources, the state needed to have a plan that indicated priorities for each project and then used administrative measures to allocate the resources accordingly to guarantee that the scarce resources would be allocated according to the state's strategic goal. In that way, market competition was suppressed.

Under such a macroeconomic-policy environment and resource-allocation system, agency problems would be a serious threat to the state's goal of maximally mobilizing resources for its priority projects. The replacement of market competition by planned allocation removed the possibility for the state to rely on the observation of relative performances to evaluate the SOE managers. Furthermore, because of the distortions in the macroeconomic-policy environment, the profitability of an SOE was determined mostly by its output and input prices. The influence of managers' actions on the profitability of an SOE was only secondary. Therefore, it was also impossible for the state to discipline the managers by simply observing the firm's profit level alone. Moreover, an incentive contract based on the comparison between current and past performances would not solve the agency

³ If only the firm's performance itself (but not the performance of other firms in the market) is observable, competition may also mitigate the agency problems by the threat of liquidation (Klaus M. Schmidt, 1997) and reduction in the firm's profits which reduce the room for managerial discretion (Oliver D. Hart, 1983), although the optimal-contracting literature suggests that these effects are not very robust.

problems because of the ratchet effect (Martin L. Weitzman, 1980) and also because the state often failed to deliver materials on time or in the quantity and quality required, so that managers could blame the state for their failures. Furthermore, it was impossible or prohibitively costly for the state to oversee the managers' actions directly. Under such a circumstance, if the state had granted business autonomy to the managers, the managers of a policy-determined profitable SOE could have a lot of shirking and on-the-job consumption because the state could observe neither their wrongdoings directly nor the firm's ought-to-be profits indirectly. For a policy-determined loss-making SOE, similar agency problems would arise because the state could observe neither the managers' discretion nor the firm's ought-to-be losses. One of the purposes of instituting price distortions in the Soviet-type economy was to maximally mobilize resources for priority projects. To prevent the policy-created economic surplus from being dissipated by managerial discretion, it was imperative for the state to deprive managers of their autonomy and to make the SOE's like puppets in the economic system. The fact that, before the recent reforms, the state always had to recentralize the management in order to control walloping increases in wages after each of the state's attempts to increase the autonomy of SOE's in China and other Soviet-type economies testifies to the above analysis.

III. The Effects of SOE Reform in China

The Soviet-type economy was very good at mobilizing resources for building a few priority sectors. However, the economy was very inefficient due to two reasons: (i) low allocative efficiency because of the deviation of the industrial structure from the pattern dictated by the comparative advantages of the economy; and (ii) low technical efficiency because the managers had no means to motivate the workers and no incentives to improve their operations (Lin et al., 1996).

To improve economic efficiency, the Chinese government initiated a series of incremental, gradual reforms in 1979 that eventually resulted in a transition to a market economy. In the process, the reforms in the

management systems led the reforms in the resource-allocation systems, which in turn led the reforms in the macroeconomic-policy environment. Specifically, the Chinese government first allowed the SOE's to share part of the performance improvement by a profit-retention program, which initially gave 12 percent of the increased profits or reduced losses to the enterprises. The SOE's could use the retained income for paying bonuses to workers, supporting welfare programs, and investing in capacity expansions. The managerial autonomy was gradually deepened through the replacement of the profit-retention system by a contract-responsibility system in which the SOE's agreed to deliver predetermined amounts of revenue to the state and retained the residuals, and later the replacement of the contract-responsibility system by the modern corporate system in which the state was entitled to the dividend on its shares in the SOE assets. Parallel to the SOE managerial reform was the decollectivization of agriculture, which replaced the production-team system with the household-responsibility system. Meanwhile, a dual-track system was introduced to reform the resource-allocation system. After fulfilling the compulsory delivery obligations, the SOE's were allowed to sell their above-quota outputs to the markets at market-determined prices. The enterprises were also permitted to purchase inputs from the markets to increase production or to expand production capacity.

An unexpected effect of the above reforms was the entry and rapid growth of nonstate enterprises, especially the township and village enterprises (TVE's). Rural industry already existed before the reform as a result of the government's 1971 policy to develop rural processing industries in order to finance the agricultural mechanization program. However, being outside the state plans, the growth of TVE's was severely constrained by their lack of access to capital, raw materials, equipment, and markets. The reforms created two favorable conditions for the rapid expansion of TVE's: (i) a new stream of surpluses were brought about by the household-responsibility-system reform and were retained in rural areas, providing a resource base for new investment initiatives (Lin, 1992);

(ii) the dual-track system provided nonstate enterprises with access to key raw materials, equipment, and markets. In 1978, the output of TVE's accounted for 7.2 percent of the total value of industrial output in China. The output share of TVE's increased to 31.1 percent in 1996.

Being outsiders to the traditional system, nonstate enterprises had to obtain credits and inputs from competitive markets, and in turn, their products were sold to markets. They faced hard budget constraints, and they would not survive if their performances were poor. The dynamism of nonstate enterprises exerted a heavy pressure on the SOE's and triggered the state's policy of deepening the SOE managerial reforms. Firm-level studies show that the increase in managerial autonomy and the intensification of competition have significantly improved the managerial incentives and total factor productivity of SOE's (Wei Li, 1997).

When the reform started in 1979, most SOE's were profitable. Taxes and revenues from SOE's were the government's main sources of fiscal income. However, in spite of the significant increase in productivity, the profitability of the SOE's has declined substantially since the reforms started. Currently, evidence shows that more than 40 percent of SOE's are operating at losses in spite of large amounts of implicit subsidies from low-interest loans and other policy protections. The decline of the profitability of SOE's is partly attributable to the dissipation of their monopoly rent. However, the walloping increases in wages and other fringe benefits are other important reasons. The average annual growth rate of the SOE wage fund in the state sector was 16 percent during 1978–1996, while the average annual growth rate of output in the same period was 7.6 percent.

IV. Policy Burdens and Soft-Budget Constraints

Before the reforms, except for shirking, agency problems such as on-the-job consumption, looting, and other wrongdoings were not serious in SOE's because of the absence of managerial autonomy. The analysis in Section II suggests that the increase of competition in a market economy should have eliminated or

at least mitigated these agency problems. However, agency problems in SOE's have worsened after the reforms in spite of the intensification of competition. What are the reasons?

From the literature, we know that the observation of some summary indicators, such as relative profits, will be a sufficient statistic of managers' actions if firms face only common uncertainties (i.e., do not have any idiosyncratic shocks in input costs, output prices, or production process) (Holmstrom, 1982). However, this condition does not hold for the SOE's in China and other transitional economies. As legacies of prereform policies, the SOE's encounter a number of idiosyncratic burdens, including the following:

- (i) The capital intensity of many large SOE's is too high, judging from the capital-scarcity nature of the Chinese economy. They cannot survive if they have to pay market-determined interest rates and face market competition, especially competition from the capital-abundant economies. Before the reforms, both their investments and working capital came from interest-free fiscal appropriations. They were also shielded from international competition. After the reform, the government replaced the fiscal appropriations with interest-bearing loans, and the protections were also gradually eliminated. In a capital-scarce economy, the capital-intensive enterprises industries are not competitive by nature. However, the state views the capital-intensive industries as strategically important and the SOE's are instructed to operate in those industries.
- (ii) The SOE's bear a heavy burden from retirement pensions, other social-welfare costs, and redundant workers. Before the reforms, the state adopted a low-nominal-wage policy. The wage was only enough to cover an employee's current consumption. The SOE's were responsible for their employees' retirement pensions, housing, medical cares, and other needs. Before the reform, this policy did not pose any extra burden on the SOE's, because the state covered all the

SOE expenditures by fiscal appropriation. However, after the managerial reform, the SOE's have had to be responsible for the wages and social welfare of not only the incumbent employees, but also their retired workers. The older an SOE is, the more retired workers it has; and the heavier the burden from retired workers' pensions and social-welfare expenditures it carries. In a similar vein, the heavy-industry-oriented development strategy before the reforms did not create enough job opportunities for urban residents. SOE's were thus forced to employ many redundant workers. In the interest of social stability, the SOE's are not allowed to lay off the redundant workers.

- (iii) Some SOE output prices are still distorted. Before the reforms, the prices of energy, raw materials, and other products or services, which were considered as inputs to the heavy-industry projects, were artificially suppressed. After 18 years of reform, most prices have been liberalized. However, the prices of energy, transportation, and a few other products are still kept below the market-equilibrium level. These prices often cannot cover production costs.

The above policy-determined burdens put the SOE's in a disadvantaged position in competing with nonstate enterprises. Because each SOE was established at a different time, has somewhat different technology and capital intensity, and has a different number of retired as well as redundant workers, the impact of the above policy burdens on the competitiveness of SOE's is idiosyncratic. Therefore, competition among the SOE's or between the SOE's and nonstate enterprises cannot serve as a device to extract information optimally. Under this circumstance, the expansion of the managerial autonomy of SOE's will worsen the agency problems.

In theory, the state should be responsible only for the SOE losses that arise from the policy burdens. However, because of the information-asymmetry problem, it is very hard for the state to distinguish between the policy-induced losses and the own operational losses

of SOE's. The managers of SOE's will ascribe all their losses to the state's policies, no matter whether the losses are due to the policy burdens or due to their own managerial discretion. Consequently, in most cases, the state in practice has to be responsible for all the SOE losses. As such, the budget constraints of SOE's become soft. The soft budget constraints in turn worsen the moral hazard, managerial slacks, on-the-job consumption, and other agency problems. To constrain the agency problems, the state will have to intervene directly into the operations of SOE's. Then there will arise a vicious cycle of policy burdens, subsidies, agency problems, and political interventions in the SOE management system.⁴

V. Fair Competition and SOE Reform

The failures of 18 years of managerial reforms to harden the budget constraints of SOE's and to improve their performance have made privatization an attractive alternative to some economists. However, as long as the policy burdens remain, even if the SOE's are privatized, the state cannot excuse itself from the policy-induced losses, and the soft budget constraints will persist. The evidence after the privatization in Eastern Europe and the former Soviet Union supports the above statement (World Bank, 1996 p. 45). Therefore, for the SOE reform to be effective, it is necessary to remove the policy burdens of SOE's and to provide them with a level playing field first. Some Pareto-improvement measures can be introduced to deal with each of the above policy burdens (Lin et al., 1998). Without policy burdens, the state is no longer accountable for failures of SOE's and can thus impose hard budget constraints on them. Without the state's subsidies, the SOE managers in turn can resist unnecessary political interventions in their

⁴ The agency literature suggests that, without competition, it is still possible to design a second-best incentive contract to minimize the principal's problem. However, the literature assumes that the shocks, no matter common or idiosyncratic, are exogenous and that the contract is enforceable. In the case of policy burdens, the state has to be accountable for the burdens. As such, incentive contracts between the state and SOE's are not enforceable and will not mitigate the principal's problem.

operations. Certainly, a level playing field does not guarantee that SOE performances will necessarily be good. If an SOE fails to perform well, other enterprises, including those privately owned, will have incentives to take over, replace its managers, improve the efficiency, and make profits from the takeover because, without policy burdens, an SOE should be able to make a normal profit with a normal management. However, whether privatization is a necessary condition for improving the efficiency of the SOE's cannot be determined a priori. Fundamentally, for any large modern corporation, as pointed out by Fama (1980), there are no owners in any meaningful sense.

REFERENCES

- Alchian, Armen A. and Demsetz, Harold. "Production, Information Costs, and Economic Organization." *American Economic Review*, December 1972, 62(5), pp. 777-95.
- Fama, Eugene F. "Agency Problems and the Theory of the Firm." *Journal of Political Economy*, April 1980, 88(2), pp. 288-307.
- Hart, Oliver D. "The Market Mechanism as an Incentive Scheme." *Bell Journal of Economics*, Autumn 1983, 14(2), pp. 366-82.
- Holmstrom, Bengt. "Moral Hazard in Teams." *Bell Journal of Economics*, Autumn 1982, 13(2), pp. 324-40.
- Li, Wei. "The Impact of Economic Reform on the Performance of Chinese State Enterprises, 1980-1989." *Journal of Political Economy*, October 1997, 105(5), pp. 1081-1106.
- Lin, Justin Yifu. "Rural Reforms and Agricultural Growth in China." *American Economic Review*, March 1992, 82(1), pp. 34-51.
- Lin, Justin Yifu; Cai, Fang and Li, Zhou. *The China miracle: Development strategy and economic reform*. Hong Kong: Chinese University Press, 1996.
- . *Sufficient information and state-owned enterprise reform*. Hong Kong: Chinese University Press, 1998 (forthcoming).
- Schmidt, Klaus M. "Managerial Incentives and Product Market Competition." *Review of Economic Studies*, April 1997, 64(2), pp. 191-213.
- Weitzman, Martin L. "The 'Ratchet Principle' and Performance Incentives." *Bell Journal of Economics*, Spring 1980, 11(1), pp. 302-8.
- World Bank. *World development report 1996: From plan to market*. New York: Oxford University Press, 1996.

China's State Enterprises: Public Goods, Externalities, and Coase

By GARY H. JEFFERSON*

This essay attempts to formulate a unified theory of the state-owned enterprise which clearly identifies the economic problem created by state ownership. The central argument of the essay is that the state-owned enterprise is a kind of impure public good with clear externality and public-policy implications. Nonexcludability and nondiminishability, properties of a public good that are inherent in the classic state-owned enterprise, create externalities that impair economy-wide economic efficiency. Remedying the externality and improving the efficiency of the state enterprise can best be achieved by applying the logic of the Coase theorem (Ronald H. Coase, 1960) to the reform of the enterprise sector in transition economies.

To examine the state enterprise as a public good, two issues require clarification. First, while pure public goods exhibit complete non-excludability and nondiminishability (or non-rivalry), in fact, few goods, not even the oceans and atmosphere, satisfy this standard. It is useful to think of most public goods along a continuum based on their relative degree of excludability and diminishability. Second, while in the past public goods were assumed to exist due to the intrinsic technical inability to limit their overconsumption, in their book on the subject, Richard Cornes and Todd Sandler (1996 p. 6) argue that "...externalities and public goods are helpfully viewed as incentive structures, rather than being inherently associated with certain activities..." The creation of an efficient incentive structure

resulting from the clear assignment of property rights and the elimination of transfer costs was the key insight of Coase (1960).

As a result of my own research focus, this essay examines the problem of the state-owned enterprise within the context of China's economy, hopefully without significant loss of generality. A fundamental characteristic of the state-owned enterprise is that it is "owned by all the people." By eroding the incentive to monitor the enterprise, this widely dispersed and ambiguous ownership structure invites the excludability problem. The state enterprise is subject to the opportunistic behavior of workers, managers, and public officials, who extract value from the firm in excess of what they put in. These exactions include asset stripping by managers, shirking by workers, predatory taxes, fees and bribes levied by public officials, and nonpecuniary benefits for employees and their relatives in the form of housing and social services. Serving a broader constituency, public officials, who treat the state enterprise as a cash cow, often direct predatory tax revenues to finance public infrastructure and services. It is this inability to monitor effectively and limit the overconsumption by large numbers of stakeholders, inside and outside the enterprise, that transforms the state enterprise into a commons.

The state enterprise is more than a commons, however. Overconsumption of the state-enterprise commons creates serious efficiency costs for the general economy. These arise from the second property of a public good. Nondiminishability within state industry means that one person's overconsumption need not seriously constrain the ability of others to extract value from the firm. With soft budget constraints, cumulative exactions that translate into losses are replenished by fiscal or financial subsidies. The asymmetry between lower jurisdictions that extract value and higher jurisdictions that replenish value, either through direct subsidies or the state banking

* Graduate School of International Economics and Finance, Brandeis University, Waltham, MA 02254. I appreciate the helpful comments of Loren Brandt, Richard Garbaccio, Albert Hu, Barry Naughton, Louis Putterman, and Thomas G. Rawski. I also appreciate the research support I received from the Henry Luce Foundation and the William Davidson Institute at the University of Michigan Business School. Errors of analysis or judgment remain my own.

system, creates a serious moral-hazard problem for opportunist local officials. These replenishments give rise to several forms of negative externality.

The first is that fiscal and financial subsidies created to replenish chronic losses require the central government to print money. The macroeconomic consequence is inflation. Alternatively, to avoid the externality of inflation, price controls may be used, but these distort resource allocations. A third form of externality is tight money and debt finance which diverts investment funds and employment-generating opportunities away from the non-state sector. A common denominator of these three forms of externality is an accumulation of bad loans and nonperforming debt within the state-enterprise sector that renders the financial system vulnerable to crisis and collapse. Externalities emerging from the state-enterprise sector bear resemblance to the air pollution that ensues from overconsumption of clean air. Overconsumption of the resource by some members of society creates externalities that affect all members of society.

The incentive problem that lies at the heart of these externalities is illustrated by Armen A. Alchian and Harold Demsetz (1972), who illustrate the importance of monitoring within a firm that depends on team production. To achieve effective monitoring, the firm requires a central contracting agent (i.e., a chief executive), who in turn is motivated by control over the residual. If the agent shirks, causing enterprise efficiency and profit to fall, the reward to the agent falls. Michael Jensen and William Meckling (1976) further observe that, in the absence of an enforcement mechanism, insiders will always want to extract a dollar value from the firm if their formal claim on the residual declines by only a fraction of a dollar, α . Where $\alpha < 1$, monitoring is needed to meter penalties for individuals whose extractions exceed their contributions. This agency problem is intrinsic to the modern capitalist publicly owned corporation. Within state ownership it is writ large, since both the authority and incentive for individual bureaucrats to curtail opportunistic overconsumption are generally lacking.

How is the problem of externality to be remedied? The formulation of the state-enterprise

problem as an externality problem invites attention to the well-known remedy prescribed by Coase (1960). According to the Coase theorem (see P. R. G. Layard and A. A. Walters, 1978, p. 192):

If costless negotiation is possible, rights are well-specified, and redistribution does not affect marginal values, then (i) the allocation of resources will be identical, whatever the allocation of legal rights, and (ii) the allocation of resources will be efficient, so there is no problem of externality. Furthermore, if a tax is imposed in such a situation, efficiency will be lost.

There are many formulations of the Coase theorem. The ambiguity arises from the first condition relating to costless negotiation, which is variously characterized as zero bargaining costs, zero negotiation costs, zero transaction costs, or even as perfect competition. While interpretations of this condition have motivated an extensive body of debate, like E. Ray Canterbury and A. Marvasti (1992), this essay broadly interprets the condition as implying a competitive market. Under this interpretation, the costless bargaining of rights between two persons is a metaphor for an efficient property-rights market just as the two-person Edgeworth Box is used to illustrate efficient exchange in the goods market. Zero transaction costs, broadly defined, imply costless information, search, entry, and exit; transparency; and ease of contracting—all key ingredients of a competitive market.

Regardless of the scope with which the first condition of the Coase theorem is interpreted (as two-person bargaining costs or a competitive market), the results of the theorem remain the same. Once a world in which property rights are not clearly assigned and exchange is costly or impossible is transformed into a Coasian property-rights market in which rights are clearly assigned and transaction costs are low, the opportunity cost of state ownership becomes well defined. It is well defined, because parties with the ability to employ assets more efficiently than existing "owners" will value them more highly and have the wherewithal to compensate existing public owners above and beyond the commercial

value of their present use. For the public owners, the question is whether the total social value of the assets in their present use is sufficient to warrant retaining the assets in light of the opportunity cost made explicit by the property-rights market.

A frequent criticism of the Coase theorem is that its conditions do not exist in practice, so that Pareto-efficient outcomes cannot be expected. The relevance of the Coase theorem to the reform of China's enterprise sector does not depend on satisfying the pure conditions of the theorem. Rather, the point is that, from a position in the late 1970's when property rights were very poorly specified and transaction costs were prohibitively high so as to preclude any form of asset exchange, the Chinese economy has moved to a stage where property rights have become sufficiently well specified and transaction costs sufficiently lower, so that exchanges are now an everyday occurrence. In important respects, the process of China's enterprise reform is a story about the assignment of enterprise property rights, the reduction of transaction costs, and the exchange of these rights among officials, managers within the firm, and outside entrepreneurs and firms in search of sales, mergers, and acquisitions. Thomas G. Rawski and I have provided an assessment of the progress of China's emerging property rights market (Jefferson and Rawski, 1997). Our results show substantial progress in some areas, such as entry, and less progress in other areas, including the exit of state-owned enterprises and transparency of township and village enterprises (TVE's). Before further examining special features of China's emerging property-rights market, I examine two other potential approaches for remedying the externality problem.

1. *Privatization.*—Some advocate that China's government immediately sell all of the state's industrial assets. The problem with this remedy is that, by itself, privatization may not effectively solve the problems of nonexcludability or nondiminishability. In the absence of a well-functioning property-rights market, privatization may result in the transfer of public assets to private agents who do not use them in a substantially more efficient way than

they had been used under state ownership. This was the case with Russia's voucherization program: insiders stripped assets knowing that newly created minority shareholders enjoyed neither the transparency nor enforcement capabilities needed to prevent this opportunistic behavior. Thus, privatization did not even result in well-specified property rights (see Maxim Boycko et al., 1993). Moreover, in the absence of hard budget constraints, insiders were able to enjoy persistent overconsumption.

2. *Hardening the budget constraint.*—Another potential remedy for the public-enterprise externality problem is to shut down the source of replenishment (i.e., to harden the budget constraint). The immediate impact of imposing a budget constraint is to transform the public good into a commons. The question then is whether making the resource rivalrous will create the incentive for stakeholders to restructure the assignment of property rights so as to curtail overconsumption. The example of fisheries in Maine suggests the potential for converting small commons into partnerships. According to James A. Wilson (1977), certain small fisheries in Maine, typically harbors, have been effectively appropriated by groups of fishermen for their sole use. By creating partnerships that prevent fishermen from other areas from "poaching," this clarification of property rights has reduced the overfishing problem.

In conclusion, privatization and hardening of the budget constraint are both elements of a Coasian remedy to public enterprise's externality problem. Although possibly the most effective method, privatization is neither a necessary nor a sufficient condition. Hardening budget constraints is a necessary but not a sufficient condition for eliminating the public-enterprise externality. For smaller enterprises, a hard budget constraint may create the incentive to make a clear and effective assignment of property rights, thus satisfying the second condition of the Coase theorem.

What are the implications of this analysis for China's township and village enterprises? It is arguable that, as public enterprises, China's TVE's are indistinguishable from

state enterprises. However, because local rural governments generally do not have the wherewithal to replenish enterprise losses continuously, TVE's are more akin to the commons, in which consumption of the resource is rivalrous. As commons with rivalrous consumption, TVE's may share a key characteristic with the small fisheries in Maine, where a self-assignment of property rights has evolved.

Consistent with this prediction, Jefferson et al. (1998) show that China's TVE's generally do enjoy a more coherent set of property rights and more effective monitoring than their state-enterprise counterparts. The proportion of TVE's that have achieved substantial reform along three measures (enterprise autonomy, concentration of internal monitoring authority, and an effective incentive structure) is, in their sample, 36.5 percent, nearly four times that of state enterprises (9.5 percent). Since these survey data were gathered during 1991–1992, an increasingly large proportion of TVE's have sought to clarify their ownership structures by converting to shareholding enterprises.

This evidence suggests that, as with the small fisheries in Maine, when faced with hardened budget constraints, stakeholders in smaller-enterprise commons are better able to self-initiate effective property-rights reform than the large numbers of stakeholders of large public enterprises, who are more likely to face insurmountable free-rider and coordination difficulties. If this is the case, then the greater incidence of hard budget constraints among China's smaller public enterprises is sensible policy since, by motivating a rationalization of property rights, hardened budget constraints create a greater payoff for these enterprises than for large public enterprises.

It is worth noting several special features of the emerging Chinese property-rights market. One such feature is the establishment of approximately 150 property-rights transaction centers throughout China. These municipal-based organizations, which bring together accounting, legal, and investment banking services and mediate thousands of sales, mergers, and acquisitions annually, represent an important component of China's emerging merger and acquisition industry (see e.g., Nicholas Howson, 1997; World Bank, 1997).

What are the implications of this Coasian market approach for the Chinese government's vision of a "socialist market economy"? A substantial portion of China's industrial assets (more than three-quarters) continues to be held by jurisdictions, public agencies, pension programs, and other forms of public ownership. According to the logic of the Coase theorem, the important condition is that whoever these owners are, public or private, their property rights should be well specified. With effective ownership of China's 1.5 million township and village enterprises and all but 1,000 or so state-owned enterprises spread over more than 46,000 subnational jurisdictions, the decentralized assignment of property rights plausibly combines public ownership with a vibrant property-rights market.

The important condition is, that by operating in the context of a property-rights market, public owners face the opportunity cost of their use of industrial assets. If public owners fail to improve monitoring and production efficiency or chose to utilize industrial enterprises to deliver social services, the persistence of economic losses should motivate an explicit accounting of the social value of public ownership. Existing public owners should always be cognizant that they have the option to engage in negotiations with other jurisdictions, enterprises, or individual entrepreneurs who can use the assets to greater advantage.

A corollary of the Coase argument is that direct government interventions, such as Pigouvian taxes on externalities, are inefficient relative to voluntary market-mediated exchanges of property rights. The Chinese government's effort to consolidate its largest state enterprises through its "1,000 Firm Reinvigoration Program" emphasizes merging strong and weak enterprises, debt forgiveness, and technical restructuring (World Bank, 1997 p. 5). Following the emergence of South Korea's financial crisis, reports indicate that China's reformers may be backing away from creating chaebol-like conglomerates through government-mandated Pigouvian "forced marriages" in favor of Coasian remedies that assign to individual enterprise directorships the right to negotiate mergers and acquisitions through the market.

China's vibrant emerging property-rights market is arguably its most valuable transition resource. Reducing restrictions on entry has caused the number of recorded industrial enterprises in China to balloon from 300,000–400,000 in 1978 (approximately the number of industrial firms in the United States) to over 7 million, representing a diverse technological and institutional mix of state, urban collective, township, village, foreign and domestic joint-venture, shareholding, cooperative, individual, and privately owned enterprises. In China and elsewhere, emerging property-rights markets perform three functions: to motivate the efficient monitoring of industrial assets to avoid the "tragedy of the commons" and costly macroeconomic externalities; to mediate the restructuring and exit of unsuccessful enterprises, both public and private; and to select sustainable forms of corporate governance that are able to withstand mounting competition within the industrial systems of transition economies. China's emerging property-rights market has begun to perform these functions.

REFERENCES

- Alchian, Armen A. and Demsetz, Harold. "Production, Information Costs, and Economic Organization." *American Economic Review*, December 1972, 62(5), pp. 777–95.
- Boycko, Maxim; Schleifer, Andrei and Vishny, Robert W. "Privatizing Russia." *Brookings Papers on Economic Activity*, 1993, (2), pp. 139–91.
- Canterbery, E. Ray and Marvasti, A. "The Coase Theorem as a Negative Externality." *Journal of Economic Issues*, December 1992, 26(4), pp. 1179–89.
- Coase, Ronald H. "The Problem of Social Cost." *Journal of Law and Economics*, October 1960, 3, pp. 1–44.
- Cornes, Richard and Sandler, Todd. *The theory of externalities, public goods and club goods*. New York: Cambridge University Press, 1996.
- Howson, Nicholas. "New Acquisition Structures in China: M&A Comes to the Middle Kingdom." Unpublished manuscript, Paul, Weiss, Rifkind, Wharton & Garrison, New York, 1997.
- Jefferson, Gary H.; Lu, Mai and Zhao, John Z. Y. "Reforming Property Rights in Chinese Industry," in G. Jefferson and I. Singh, eds., *Reform, ownership, and performance in Chinese industry*. New York: Oxford University Press, 1998 (forthcoming).
- Jefferson, Gary H. and Rawski, Thomas G. "Chinese Enterprise Reform as a Market Process." William Davidson Institute Working Paper No. 76, University of Michigan Business School, June 1997.
- Jensen, Michael and Meckling, William. "Theory of the Firm: Managerial Behavior, Agency Costs, and Ownership Structure." *Journal of Financial Economics*, October 1976, 3(4), pp. 306–60.
- Layard, P. R. G. and Walters, A. A. *Microeconomic theory*. London: McGraw-Hill, 1978.
- Wilson, James A. "A Test of the Tragedy of the Commons," in Garrett Hardin and John Baden, eds., *Managing the commons*. San Francisco, CA: Freeman, 1977, pp. 96–111.
- World Bank. *China's management of enterprise assets: The state as shareholder*. Washington, DC: World Bank, 1997.

Village Leaders and Land-Rights Formation in China

By SCOTT ROZELLE AND GUO LI*

In the debate about land rights in China, almost none of the participants discusses the great range of rights found throughout the country. Some argue that most village leaders take all land back periodically and reallocate the plots among the households (Guangzhou Wen, 1995; Roy Prosterman et al., 1996). Others believe that land reallocation occurs because of demographic change, and that reallocation mainly equalizes the land-to-man ratios among farm households (Qiren Zhou, 1994).

Most participants also ignore the issue of land-rights formation. Their works either imply that land rights have arisen due to forces outside the village or simply leave unexplained the question of where rights have originated. For example, in the view of some, the central government controls rights arrangements, and village leaders, acting primarily as representatives of the state, carry out the policies handed down to them (Xiaoyuan Dong, 1995; Prosterman et al., 1996). Others imply that village leaders' rent-seeking activities may play a crucial role in how land is managed (D. Gale Johnson, 1995). A third group implies that current arrangements arise from farmers' demands and are consistent with their interests (James Kung and Shouying Liu, 1996).

This paper argues that these common perceptions of the homogeneity of land rights and the existence of some single, exogenous determining force in land-rights formation either are incomplete or inconsistent with actual observations in the field. The lack of rigorous theoretical and empirical research on land-rights formation often has resulted in misperceptions about how to evaluate the current rights arrangements. This paper argues that

village leaders play an important role in land-rights creation, yet the land-rights literature almost never discusses their behavior. The relationship between these local leaders and farmers, and between local leaders and their administrative superiors, may affect leader behavior and in part influence how land rights are formed.

I. Land Rights in China

Analyzing land management by using traditional tenure types (e.g., private plots or responsibility land) may be misleading because each form represents a bundle of rights, and that bundle may vary over time and from place to place. In the case of responsibility land, for example, in 34 percent of the sample villages, farm households have secure tenure rights; that is, farmers still farm the same cultivated area they farmed in the early 1980's. Leaders in these villages have not adjusted the distribution of responsibility land for more than 15 years, even though this practice is legally sanctioned by national leaders. Large variations in the frequency of land readjustments also appear. In 36 percent of the villages, land readjustment took place only once. About 26 percent of village leaders have redistributed land twice since the early 1980's. In the remaining 38 percent of villages, land readjustment occurred three times or more.

While central government policy allows for multiple tenure types, it does not clarify the level at which land-rights decisions are to be made. If central policymakers play a dominant role, a fairly homogeneous land-rights system should exist among every village. If local leaders play an important role, however, there may be large differences among villages, even within a single township. Differences in the degree of land security exist not only among different provinces, but also from township to township within a county and from village to village within a township (Li and Rozelle, 1997). In terms of land readjustments, in 39

* Department of Agricultural and Resource Economics, University of California, Davis, CA 95616, and Food Research Institute, Stanford University, Stanford, CA 94305, respectively.

out of 44 sample counties (87 percent), two townships within a county reported different land-readjustment frequencies. In 52 out of 92 townships (57 percent), two villages within a township reported different land-readjustment frequencies. The observed heterogeneity in fundamental land-management practices essentially weakens the assumption that land rights are the same across China and bolsters the proposition that real land-rights formation occurs at the village level.

While land rights vary even among villages within provinces, counties, and townships, they are not random. The sample villages can be divided into two types: villages that have readjusted land since implementation of the household-responsibility system (HRS), and villages that have not. Village leaders in villages with more interest to protect appear systematically to control land more frequently than leaders in more remote, poorer villages. For example, the size of the village fund is larger (880 yuan vs. 340, statistically significant at the 5-percent level) and the percentage of villages with village-run industries is greater (35 percent vs. less than 25 percent) in villages where leaders have adjusted land than in those that never have. However, leaders in richer, more urban areas control land less frequently.

Leaders in regions where land is more "valuable" also control land more often than leaders in villages where land is less scarce (Li and Rozelle, 1997). In villages that have readjusted land since HRS, there is only 0.17 hectare per capita, the population growth rate is 5.1 percent, and the proportion of irrigated land is 56 percent. In contrast, in villages that have not readjusted land, there is 0.20 hectare per capita, the population growth rate is 3.7 percent, and the proportion of irrigated land is 52 percent. Hence, leaders have a greater propensity to adjust land when there is less land, a higher population growth rate, and land of higher quality.

Finally, leaders in villages that are more rural (and by assumption have poorer on-farm labor markets and land markets [Li and Rozelle, 1997]) and those in villages that have considerable movement of workers into the off-farm labor markets (so land is less equitably distributed) are more likely to adjust the holdings of their farmers.

II. Village Leaders and Land: A Conceptual Framework

The challenge for those interested in explaining land rights formation is to come up with a theory that explains why local leaders set rights in the way that they do. The theory needs to account for the observed facts and be consistent with the political-economy realities found in the field. This section sketches our explanation, delineates the implied hypotheses, and derives the empirical model to test the hypotheses implied by the model.

Li and Rozelle (1997) show that village leaders, in pursuit of their objectives (and subject to local policy and endowment constraints), use resources under their control to induce villagers to behave in ways that are consistent with the leaders' objectives. Among local factors, land may be the most important resource used by leaders to influence villager behavior, since in many places, it may be the only resource leaders can use to influence villagers' actions.

Given such a relationship between leaders and villagers and the central role played by land, understanding how leaders use land to further their goals becomes an important step in analyzing the determinants of land rights. This section briefly describes the work by Li and Rozelle (1997), which offers an explanation of how leaders may be making decisions on land tenure-security rights. The section outlines three fundamental motivations (interest protection, administration-cost minimization, and pursuit of equity) that guide land-rights determination and delineates a series of hypotheses that would be true if the proposed motivations were accurate.

A. Interest Protection

Interest-protection behavior can establish a direct link between important behavioral objectives of leaders and land-rights determination. The level of a village leader's personal income and his status in the local community are dependent on whether or not the leader can retain his position as a village leader. If the leader loses his position, his personal income may decline as he relinquishes control over certain resources, for example, village enterprise management or control of the village treasury. To

keep his position, the village leader must fulfill a number of administrative tasks, especially "hard" policy constraints such as family-planning targets, grain quotas, corvée labor obligations, and different kinds of taxes,—imposed upon him by higher levels of government. The village leader can use his de facto control over land resources, a scarce resource demanded by farmers, to accomplish these tasks by designing a package of rights that induces a certain kind of behavior among farmers and maximizes the likelihood that farmers carry out their administrative duties. In terms of tenure security, for example, the leader can take land away from farmers if they are not satisfying village-assigned tasks, or give farmers more land when they do. The process whereby the leader uses land as a carrot or stick to get farmers to fulfill local policy goals is called "interest protection." The leader's own interests are protected because satisfied higher-level officials allow the local leader the latitude to pursue his own economic goals.

B. *Minimizing Administrative Costs*

In most places, certain administrative tasks, such as collecting compulsory grain-procurement quotas or enforcing birth control, may take a great deal of time to accomplish. Leaders always complain that they spend a large amount of time making sure that the state's demands are satisfied. Without endangering their positions, however, leaders want to spend as little of their limited time as possible carrying out administrative duties, in order to leave as much time as possible to pursue income-earning activities and leisure. The process by which leaders accomplish their duties in as little time (or at as little cost) as possible is called "effort minimization." Leaders in some circumstances use land to push farmers to fulfill their administrative tasks if the benefit of doing so (i.e., the reduction in enforcement effort) at least equals the administrative cost of implementing land-rights arrangements.

C. *Improving Equity and Production Efficiency*

Village leaders also may be trying to improve equity and production efficiency when

they make decisions about land rights. As caretakers of land, one of China's scarcest resources, leaders naturally are concerned about its efficient use. Equity is important to both villagers and village leaders (Rozelle, 1994). Leaders themselves are local residents and have close kinship and friendship ties with villagers. Eliminating inequality can boost their reputations as good headmen because villagers dislike inequality. Moreover, rich-poor gaps may make it harder to fulfill administrative tasks, imperiling leaders' status, promotion possibilities, and job security. In regions where disequilibrium exists in land distribution, and where there are no land or labor markets for equalizing resource ratios, leaders may have an incentive to reallocate land.

D. *Implications for Land-Rights Formation*

If leaders make decisions about land organization based on the above three motivations, four sets of hypotheses can be derived and subjected to empirical testing as a way to check the consistency of the framework with reality.

Hypothesis 1: In villages with profitable village enterprises or a large village fund, there is a higher likelihood that leaders will control land rights, because they have greater interests to protect.

Hypothesis 2: In villages where villagers have high-income earning opportunities, there is a lower likelihood that leaders will control land, because the loss that the leaders would incur if removed from office would be lower.

Hypothesis 3: In villages where land is more valuable (i.e., in those places with low land per capita, high population growth rates, and more irrigated area), leaders are more likely to control land rights since it will take less effort to "persuade" farmers to meet administrative targets.

Hypothesis 4: In rural villages that are more remote from urban centers (since these areas are most likely to have poor land and on-farm labor markets), and where villagers have different off-farm job opportunities, there is a

higher likelihood that village leaders will exercise control over land. In such villages, leaders can improve equity and efficiency by equalizing marginal products of land and labor across households.

III. Model and Data

To test these hypotheses, a model describing land-rights formation is specified as a function of village-leader interests, opportunity income, stringency of the administrative tasks, the "value" of the land in farm income, and the nature of the land and labor markets. The empirical model is described in detail in Li and Rozelle (1997).

The data used for the study is from a village-level community survey conducted by the authors in the summer of 1996. The survey covered 184 randomly selected villages in six representative provinces across China. In the section on land management, village leaders answered detailed questions about land-readjustment activities. The main measure of tenure security is derived by assigning a dichotomous variable the value 1 if the village had never readjusted land between reform and 1995, and 0 if leaders had readjusted land. An alternative measure was also developed using the frequency of land readjustment. The data set also includes a number of other variables measuring the interests of leaders (e.g., the existence of village enterprises; the size of the village welfare fund), the opportunity cost of leaders (village per capita income), the value of land (cultivated area per capita; growth of village population; proportion of cultivated land that is irrigated), the stringency of the administrative task (the size of the quota), the completeness of markets (measures of urbanization), and the degree of labor in the off-farm labor market.

IV. Results

The analysis used a probit specification, a standard nonlinear discrete-choice model, for the leader's discrete land-rights decisions (first column in Table 1). For the purpose of comparison, the analysis also uses a linear probability specification (see Li and Rozelle, 1997). To avoid simultaneity and heterogeneity bias,

TABLE 1—RESULTS OF ESTIMATION OF SECURE LAND-TENURE RIGHTS

Independent variables	Dependent variable	
	Discrete choice ^a	Frequency ^b
Interest protection		
Village administrative expenditure	-0.008 (-1.09)	-0.36 (-0.24)
Village enterprises dummy (1 yes, 0 no)	-0.130 [†] (-1.68)	-0.59* (-2.41)
Urbanization index	0.271** (2.88)	0.33 [†] (1.93)
Income per capita	0.087* (1.98)	1.67 [†] (1.74)
Minimizing enforcement cost		
Population growth rate	-0.006 [†] (-1.63)	0.39 (0.48)
Proportion of land irrigated	-0.001 (-0.96)	-0.30 (-1.02)
Urbanization index × income per capita	-0.026* (-2.18)	-0.44 [†] (-1.82)
Land per capita	0.084** (2.80)	0.18** (2.68)
Proportion of off-farm labor	0.016 [†] (1.64)	2.38 (1.36)
Quota	0.007 (1.38)	-0.27 (-0.27)
Improving both equity and efficiency		
Urbanization index × proportion of off-farm labor	-0.048 [†] (-1.75)	-0.53 (-1.24)
Pseudo R ² :	0.45	0.45

Notes: Numbers in parenthesis are *t* ratios. Coefficients of constant, provincial dummies and control variables are not presented (see Li and Rozelle, 1997).

^a Value equals 1 means the village has never readjusted land since HRS was introduced in early 1980's, otherwise, equals 0. Probit estimation.

^b OLS estimation. Frequencies of readjustments multiplied by -1, so predicted signs on coefficients in both columns will be the same.

[†] Statistically significant at the 10-percent level.

* Statistically significant at the 5-percent level.

** Statistically significant at the 1-percent level.

the village leader's decision on land rights in 1995 are explained by independent variables measured in 1988 and a set of provincial dummy variables. An ordinary least-squares (OLS) estimator was used to measure the impact of leader, village, and other characteris-

tics on the frequency of adjustment (last column in Table 1), and the results from the two models are fairly consistent.

As predicted by the model, leaders' interests have important effects on their choice of land rights. With more interests to protect, the probability that leaders readjust land increases. For example, when leaders are running village-owned enterprises and have access to the firm's earnings and control of its assets, the likelihood that leaders readjust land and the frequency of readjustments increase (second row in Table 1). Although the *t* ratios are small, the signs on the village-welfare-fund variable also are consistent with the predicted hypothesis (first row). When there are assets to control, there is less of an inclination to provide farmers with secure land rights.

In contrast, if leaders have high alternative income-earning opportunities, the propensity to readjust land falls. For example, when income per capita in the village is higher, the probability that leaders readjust land decreases by 8.7 percent (and the coefficient is significant at the 5-percent level [fourth row]). When villages are located in urban areas (where it is assumed that there are many income-earning opportunities inside of the village), leaders also are less inclined to spend their effort trying to control land rights (third row). In general, then, the results are consistent with predictions: the more personal interests that need protection, the more likely it is that leaders will control land rights.

The results from the estimation also support the predictions that leaders choose to control land rights if it reduces the cost of carrying out their administrative tasks (rows 5–9). Since it is easier to use land as a reward or punishment when there is a high demand for land, the probability of controlling land rights rises as the value of land increases. For example, as the village's population growth rate rises, and the land-to-person rates get smaller, the probability that leaders will readjust land will increase. Also, when farmers have access to off-farm activities, the relative value of land is lower, and leaders find it less attractive to readjust land.

The results also show that when on-farm labor and land rental markets are incomplete, leaders may use land readjustments to improve

both the equity and efficiency of farmers in villages where access to off-farm jobs may have created an imbalance among villages in their landholdings. In villages with a greater proportion of villagers working off the farm and in villages that are more remote from urban centers (assuming that labor and land markets are poorer in these remote areas), the probability that leaders will readjust land increases about 4.8 percent (and is significant at the 10-percent level [last row]). The results for a similar set of regression on land-transfer rights, provide similar support for the hypotheses (Li, 1997).

V. Conclusion

Based on a national representative data set, this paper demonstrates that land-rights heterogeneity characterizes China's agricultural economy. The paper provides evidence that land-rights variations among villages are due to systematic differences in the way local authorities manage land resources. These findings fundamentally invalidate the existing common perceptions of the homogeneity of land rights in China.

This paper also offers an innovative explanation of land-rights formation in China. Based on field observations, descriptive statistics, and the empirical results, land rights may be set by local leaders who are pursuing three objectives: (i) protection of the leader's personal interest; (ii) minimization of the administrative costs; and (iii) improvement of both equity and land-use efficiency in remote areas characterized by subsistence agriculture and unequal access to off-farm labor activities. If the central leaders better understand the motives of those who make decisions on land rights, they may be able to design better policies.

REFERENCES

- Dong, Xiaoyuan. "Two-Tier Land System and Sustained Economic Growth in Post-1978 Rural China." Working paper, University of Manitoba, 1995.
- Johnson, D. Gale. "Property Rights in Rural China." Working paper, University of Chicago, 1995.

- Kung, James and Liu, Shouying. "Land Tenure Systems in Post-Reform Rural China: A Tale of Six Counties." Working paper, University of Hong Kong Science and Technology, 1996.
- Li, Guo and Rozelle, Scott. "Land Rights, Tenure, and Leaders in China." Working paper, Food Research Institute, Stanford University, 1997.
- Prosterman, Roy; Hanstad, Tim and Ping, Li. "Can China Feed Itself?" *Scientific American*, November 1996, 275(5), pp. 90-96.
- Rozelle, Scott. "Decision-Making in China's Rural Economy: The Linkages Between Village Leaders and Farm Households." *China Quarterly*, March 1994, (137), pp. 99-124.
- Wen, Guangzhou. "The Land Tenure System and Its Saving and Investment Mechanism: The Case of Modern China." *Asian Economic Journal*, November 1995, 9(3), pp. 233-59.
- Zhou, Qiren. "Land System in Rural China: The Case in Meitan County of Guizhou Province," in G. J. Wen, ed., *The land system in contemporary China*. Changsha, China: Hunan Science and Technology Press, 1994, pp. 37-104.

BANKING CRISES, CURRENCY CRISES, AND MACROECONOMIC UNCERTAINTY

The Double Drain with a Cross-Border Twist: More on the Relationship Between Banking and Currency Crises

By VICTORIA MILLER *

Southeast Asia has recently been embattled by a plague of currency and banking crises. Thailand, Indonesia, Malaysia, and the Philippines have all experienced both of these types of crises during the last year. However, the joint occurrence of these two crises is not restricted to Southeast Asia: several Latin American countries experienced both during the last two decades (Chile in 1982, Argentina in 1982 and 1995, Venezuela in 1994, and Mexico during 1994–1995 to name only a few); the same was true of Finland, Norway, and Sweden at the beginning of the 1990's; and there is a reasonable probability that by the time the century is up, the three Baltic States will have also experienced both banking and balance-of-payments problems.

Given the recent epidemic of banking and balance-of-payments crises, researchers have started considering how these two types of crises may be related. The conclusion of that research is that causation may run in either direction and that there is an important complementarity between bank solvency and currency stability.

Maurice Obstfeld (1994) argues that a weak banking sector may itself precipitate a currency crisis if rational speculators anticipate that policymakers will not choose to endure the costs of defending their currency. Andres Velasco (1987) and Guillermo Calvo (1995) also show that an internal drain (i.e., bank run) can cause an external drain (i.e., speculative

attack on a currency) if the increased liquidity which results from a government bailout is inconsistent with the fixed parity. In Miller (1997) I follow this same line of reasoning and explicitly consider the policy alternatives available to a government that is confronted by bank runs in a fixed-exchange-rate regime.¹ Finally, Brenda Gonzalez-Hermosillo (1996) shows that the same direction of causation will result if the financial system is poorly developed and agents substitute foreign assets for domestic deposits.

While Graciela L. Kaminsky and Carmen M. Reinhart (1995) find empirically that bank crises have preceded many of the currency crises that have occurred over the last two decades, causation could still run in the other direction. For example, Miller (1996a) demonstrates that a speculative attack on the currency can give rise to a banking crisis if deposit money is used to speculate on the currency and banks are "loaned-up."² Rojas-Suarez and Weisbrod (1995) and Obstfeld (1994) also discuss how a currency crisis can create problems for a vulnerable banking sector if the government defends its currency and increases interest rates.

In the present paper I continue the study of the possible linkages between currency and banking crises. However unlike the preceding studies, the present text gives the

* Département des Sciences Économiques, Université du Québec à Montréal, Case Postale 8888, Succursale A, Montréal, Québec H3C 3P8. I thank Luc Vallée for his, as always, great comments.

¹ Liliana Rojas-Suarez and Steven R. Weisbrod (1995) also address the choice between banks and the currency.

² Some authors argue that such causation occurred in the United States in 1893. See Miller (1996b) for a complete discussion of the United States' currency and banking crises during that time.

"double-drain" literature a cross-border twist by providing examples of how a banking (currency) crisis in one country can give rise to a currency (banking) crisis in another country. Thus, I integrate the literature on the relationship between currency and banking crises with that of the international transmission of financial crises.

The transmission of financial crises across international borders has become increasingly relevant over the last few years. One need only consider the magnitude of the 1994 "tequila effect" (which hit as far east as Singapore!) and the number of Southeast Asian currencies that followed the Thai baht's plunge to realize that crises now spread like wildfire. Moreover, increased globalization and capital flows mean that industrial countries are no longer exempt. For example, given the heavy exposure of Japanese banks to Southeast Asia,³ the economic turmoil in that region could further weaken Japan's already ailing banking sector.⁴ Such a deterioration would likely be transmitted around the world. As Joe Peek and Eric S. Rosengren (1997) recently demonstrated, the 1989–1992 Japanese stock-market decline was transmitted to the United States via a drop in the lending activities of U.S. branches of Japanese parent banks. The recent international shock waves that followed the dive of Hong Kong's Hang Seng only lends further support to the fact that financial stability at home has never depended more on what is happening abroad.

I. From Bank Runs at Home to Currency Crises Abroad

Below, I provide two examples of this direction of causation. In the first, banks are the principal conduit of capital across international borders. In the second, banks lend to domestic agents while individual depositors and investors import and export capital. In

both examples, the domestic economy is large, and the foreign one is small. Moreover, the foreign country pegs the value of its currency to the national money, and there is no deposit insurance or lender of last resort.

Example 1: Domestic banks are important extenders of credit to the foreign country.

Suppose that there are two periods, that depositors invest only in domestic banks, and that banks invest at home and abroad so that interest parity is satisfied. Moreover, in a non-panic state of the world, utility-maximization implies that consumption is positive in each period. Thus, depositors withdraw a positive amount each period, and banks plan their investments so that expected inflows and outflows are equal.

It is assumed that, in the absence of runs, the amount of money that banks repatriate from abroad each period does not strain the foreign central bank's foreign-exchange reserves in any significant way. However, if all foreign investments are repatriated in a single period, then those reserves will be depleted, and the foreign central bank will be forced to devalue its currency.

Given the no-run equilibrium, depositors can self-generate a rational bank run simply by believing that other depositors will demand the entire value of their claims in the first period. To see this, note that banks initially make their investments so that inflows equal outflows, which are strictly positive in both periods of the no-run and thus no-devaluation state of the world. As banks will be unable to repatriate all of their foreign investments in the first period before the foreign currency is devalued, the devaluation will reduce the domestic-currency value of those investments and cause banks' assets to become worth less than their liabilities. In other words, banks will become insolvent. As banks will expect such a loss in the wake of a run, they will rush to repatriate the entire value of their foreign investments. This will materialize as a speculative attack on the foreign central bank which is staged by domestic banks. The self-generating nature of the domestic bank run and speculative attack on the foreign currency are formally illustrated in Miller (1998).

³ The exposure is direct through cross-border lending, and indirect through loans to Japanese companies that operate in the region.

⁴ U.S. exports could observe a similar decline if the capital flows that finance the part of those exports that are destined for Latin America run dry.

The above scenario highlights a potential source of further instability in Southeast Asia which stems from Japan's ailing banking system. Spectacular growth rates in Southeast Asia over the last decade led to a boom in Japanese lending to the region: more than half of the external debt of Thailand and a third of Indonesia's is owed to Japanese banks. While Japanese banks presently seem to be committed to staying in the region, further financial weakness in Japan threatens a repatriation of capital and further declines in the external values of Southeast Asia's currencies. The exact magnitude of Japanese banking difficulties that would lead to a transmission to Southeast Asia and the conditions under which such a transmission would occur certainly warrant further investigation.

Example 2: Private investors invest abroad.

When banks face a run and there is no lender of last resort, cash payments are typically suspended or restricted until the panic subsides or banks are able to get enough liquidity to pay off frightened depositors. When such restrictions occur, as was the case in the United States before the Federal Reserve System was established in 1934, currency (as well as foreign exchange) often sells at a premium to domestic deposits.⁵ As a currency premium increases the expected return on domestic deposits, it encourages capital inflows. Thus, a bank run in a large country that results in a currency premium can attract so much capital from a small foreign country that the small country's foreign-exchange reserves become exhausted, and it is forced to devalue.

To see why a currency premium encourages capital imports let ρ denote the domestic currency premium. During the premium period, investors could take one dollar of domestic money, convert it into $1 + \rho$ dollars of deposits, and then withdraw this amount plus interest without penalty after the restriction is lifted. A foreigner who buys domestic bank deposits and sells them after the restriction is

lifted would earn a return of $i + \rho + x$ where i is the domestic interest rate and x is the rate of depreciation of the foreign currency during the premium period. Thus, if $i + \rho + x$ is (expected to be) greater than the return on foreign deposits, then capital will flow into the domestic economy.

To get an idea of how large such capital flows can be, consider the experience of the United States in 1893. In the late fall of that year, bank runs culminated in a cash-payments restriction and currency premium. As the restriction was not expected to last for very long, it gave rise to a dramatic importation of gold even though there is evidence that agents were speculating against the dollar just before (see Miller, 1996b). Indeed, gold flows were so great at the time that during the four weeks of the premium, \$40 million of gold entered the United States, which was about half of the treasury's free gold reserves. Given the magnitude of capital flows today and the extent of financial integration worldwide, the eagerness of investors to capitalize on a foreign-currency premium could easily result in a depletion of a small country's foreign-exchange reserves and a devaluation of its currency. This did not happen in 1893, however, because under the classical gold standard of the time, most countries' monetary bases were fully backed by gold. Today, however, as liabilities of consolidated banking systems are not completely backed by foreign exchange, foreign-exchange reserves can become exhausted before the demand for them is satisfied. When this occurs, a currency is devalued.

II. From Currency Crises Abroad to Banking Crises at Home

Example 3: Domestic banks lend to domestic companies that are highly exposed abroad.

Consider a country in which banks lend to domestic companies that export to a foreign country. A devaluation of the foreign currency could then significantly damage firms' abilities to repay their loans and thus harm domestic banks by reducing competitiveness. A similar sequence of events will occur if firms borrow at home but operate abroad and if the currency crisis abroad interrupts economic activity and thus the local demand for goods.

⁵ Currency premia have been observed during all cash-payment restrictions that occurred in the United States before the creation of the Federal Reserve System.

The above example again brings to mind the recent devaluations of Southeast Asian currencies, as a decline in Japanese exports could further strain Japan's already weak financial sector. While only about 20 percent of Japan's exports go to Southeast Asia, about a third of all foreign direct investment in the region comes from Japan; and 70 percent of Japanese production in the region is used to meet local demand. Thus, it is easy to understand why Southeast Asia's currency crisis poses a serious threat to Japanese banks. Again, the conditions under which financial crises are transmitted across international borders must be investigated further to determine, for example, whether an IMF-type plan should be put in place in order to prevent further hemorrhages of capital from Southeast Asia from causing a full-blown banking crisis in Japan.

Example 4: Currency mismatches.

A currency crisis in a foreign country could also cause a banking crisis at home if, as is illustrated in Miller (1998), bank portfolios are "currency-mismatched" in the sense that more assets than liabilities are denominated in terms of a devaluing foreign currency. A similar outcome could result if firm revenues are pegged to the foreign currency while loans are denominated in the domestic one. In such a scenario the devaluation of the foreign currency would strain companies' abilities to repay domestic bank loans.

While American and European banks follow strict guidelines in managing their currency exposures, the same does not appear to be the case in emerging markets. The failure to comply with prudential rules regarding currency exposures makes emerging countries' banking systems vulnerable to currency fluctuations. For example, the currency mismatch of Thai banks resulted in a double drain on domestic soil last year;⁶ and similar apparent imbalances in Estonia make its banking sys-

tem vulnerable to movements in the deutsche mark.⁷

III. Conclusion

I have argued that, while a banking (currency) crisis may cause a currency (banking) crisis within a given country, such transmission mechanisms can also operate across international borders. I have provided a few examples of how this can be the case. Those examples indicate that, just as creditors consider the financial health of potential borrowers when deciding whether or not to lend money, borrowers should also evaluate the health of potential creditors when deciding from whom they will borrow. However, as the smooth functioning of a country's financial system depends increasingly on financial stability abroad, further research on the subject is required so that policymakers can efficiently engage in the prevention of crises and their transmission internationally.

The cross-border externalities of financial-sector weaknesses and the increasing need for financial information from abroad, suggest that supervision by an international agency such as the IMF should be considered as a tool for crisis prevention and containment (just as intraborder externalities justify supervision by a domestic government). However, while the case for an international regulator seems easy to make, the case for an international lender of last resort is less evident. While such a lender would be desirable in order to prevent the spreading of financial crises across international borders and to minimize their fallout at home, one would want to avoid the irresponsible policies that could follow expected bailouts.

⁶ Thai banks took advantage of lower foreign interest rates to borrow in foreign currencies and to lend in the domestic one. As they did not cover these positions, the devaluation of the baht seriously compromised the solvency of those banks.

⁷ Estonia's currency, the kroon, is pegged to the deutsche mark, while 80 percent of its banking system's loans are denominated in foreign currencies. While it is difficult to know the percentage of those loans that are in currencies other than marks, if it is a significant percentage then an appreciation of the mark will seriously compromise the solvency of that banking system. A devaluation of the kroon against the deutsche mark however, would provide a capital gain to banks.

REFERENCES

- Calvo, Guillermo. "Varieties of Capital-Market Crises." Working paper, University of Maryland Center of International Studies, 1995.
- Gonzalez-Hermosillo, Brenda. "Banking Sector Fragility and Systemic Sources of Fragility." International Monetary Fund Working Paper No. 96/12, February 1996.
- Kaminsky, Graciela L. and Reinhart, Carmen M. "The Twin Crises: The Causes of Banking and Balance-of-Payments Problems." Mimeo, International Monetary Fund, Washington, DC, 1995.
- Miller, Victoria. "Speculative Currency Attacks with Endogenously Induced Commercial Bank Crises." *Journal of International Money and Finance*, June 1996a, 15(3), pp. 383-403.
- . "Exchange Rate Crises with Domestic Bank Runs: Evidence from the 1890s." *Journal of International Money and Finance*, August 1996b, 15(4), pp. 637-56.
- . "Central Bank Reactions to Bank Crises in Fixed Exchange Rate Regimes." Mimeo, Université du Québec à Montréal, 1997.
- . "Domestic Bank Runs and Speculative Attacks on Foreign Currencies." *Journal of International Money and Finance*, 1998 (forthcoming).
- Obstfeld, Maurice. "The Logic of Currency Crises." National Bureau of Economic Research (Cambridge, MA) Working Paper No. 4640, February 1994.
- Peek, Joe and Rosengren, Eric S. "The International Transmission of Financial Shocks: The Case of Japan." *American Economic Review*, September 1997, 87(4), pp. 495-505.
- Rojas-Suarez, Liliana and Weisbrod, Steven R. *Financial market fragilities in Latin America: The 1980s and 1990s*. International Monetary Fund (Washington, DC) Occasional Paper No. 132, October 1995.
- Velasco, Andres. "Financial Crises and Balance of Payments Crises: A Simple Model of the Southern Cone Experience." *Journal of Development Economics*, October 1987, 27(1-2), pp. 263-83.

Financial Crises in Asia and Latin America: Then and Now

By GRACIELA L. KAMINSKY AND CARMEN M. REINHART*

The devaluation of the Thai baht on 2 July 1997 generated waves of turbulence in currency and equity markets that surpassed the "tequila" effects in the wake of the 1994 devaluation of the Mexican peso. The crisis first spread to East Asia in the form of a string of devaluations and stock-market collapses. As the problems intensified, the currencies of other Asian countries, including Hong Kong and South Korea, came under speculative pressure. Outside the region, Argentina, Brazil, and Russia suffered sharp declines in their equity markets and periodic bouts of speculation against their currencies. As the dust settles in currency markets, many of these countries will be left with serious banking-sector problems, if not full-scale banking crises, as in Thailand and South Korea.

Our earlier work on financial crises suggested that economies behave differently on the eve of crises (see Kaminsky and Reinhart, 1996). Typically, financial crises occur as an economy enters a recession that follows a prolonged boom in economic activity fueled by credit creation and surges in capital inflows. The cycle of overlending is exacerbated by implicit or explicit deposit guarantees, poor supervision, and moral-hazard problems in the banking sector.¹ Crises are accompanied by an overvaluation of the currency, weakening exports, and the bursting of asset price bubbles.

In this paper, we extend that work by analyzing the extent to which past crises share common characteristics in Latin America, Asia, Europe, and the Middle East. In addition, we examine the recent crises in Asia and in Latin America to determine whether the con-

siderable regional differences that we find for the earlier sample have eroded.

I. Regional Differences

Between 1970 and 1995, Latin American countries (LA) suffered 50-percent more crises (per country) than the East Asian countries (EA), or the European and Middle Eastern countries (Others) in our sample.² This section examines whether currency and banking crises differed across regions.

We begin by examining the behavior of 15 economic indicators. The indicators used to capture the overlending cycles include the M2 multiplier, the ratio of domestic credit to nominal GDP, the real interest rate on deposits, and the ratio of lending-to-deposit interest rates. Increases in any of these indicators might signal possible financial-sector problems.³ Other financial indicators include "excess" real M1 balances (to capture lax monetary policy), deposits at commercial banks (to assess whether there are runs in the midst of the crises), and the ratio of M2 (in dollars) to foreign-exchange reserves in dollars (to examine to what extent the liabilities of the banking system were backed by international reserves). The current-account indicators are exports, imports, terms of trade, and deviations of the real exchange rate from trend. Declines in exports and the terms of trade, increases in imports, and real appreciations of the domestic currency signal potential problems in the current account. The capital-account indicators are foreign-exchange reserves of the central bank and domestic-foreign real-interest-rate differentials. Reserves losses and increasing interest-rate differentials are signals of future

* Board of Governors of the Federal Reserve System, Washington, DC 20551, and School of Public Affairs, University of Maryland, College Park, MD 20742, respectively. The views expressed in this paper are those of the authors and do not necessarily reflect those of the organizations with which they are affiliated.

¹ See Ronald McKinnon and Huw Pill (1994) on the interaction of capital inflows and liberalization.

² For a discussion of how the crises are defined and dated, and countries in the sample, see Kaminsky and Reinhart (1996).

³ See Hali J. Edison and Marcus Miller (1997) on credit cycles.

problems in the capital account. Finally, we include output and stock prices (in dollars), with declines in output and stock-market crashes signaling impending crises. The interest rate, the spreads, "excess" real balances, and real exchange rate deviations from trend are in levels, and all other indicators are 12-month percentage changes.

In an earlier paper, we concluded that the behavior of most of these indicators in the months prior to the crises departed significantly from the behavior in "tranquil times," which is defined as all the months in the sample outside the 36 months around the crises. For instance, there were unusually large declines in equity prices relative to tranquil periods on the eve of the financial crises. To assess whether these pre-crisis deviations in individual indicators are larger in LA than in other regions, we measure *volatility* by calculating the mean absolute deviation from tranquil periods (as percentages) for each indicator in the 18 months prior to the crisis for the three regions separately. The first column in Table 1 lists the indicators; the second column reports the mean deviation from tranquil periods for that indicator for LA; the third and fourth columns report the comparable mean absolute deviations for EA and Others, respectively. An asterisk denotes that the regional difference from LA, which is the benchmark, is statistically significant at the 5-percent confidence level. If we compare EA and LA currency crises (Table 1A) 10 of the 15 indicators are significantly more volatile for LA, including all the financial and capital-account indicators. The regional patterns for banking crises paint a similar picture, as the amplitude of the pre-crisis cycles relative to tranquil times is larger for LA than elsewhere. However, regional differences in volatility are diminishing. For instance, the decline in Thai equity prices (in U.S. dollars) since their 1995 peak exceeds 80 percent; the magnitude of this deviation from the norm during tranquil periods is more in line with those observed in the LA crises than in previous EA crises. Similar anomalies are evident in other indicators, such as credit.

While Table 1 presents evidence on the volatility of individual indicators, we now focus on their behavior as a group. As to the *fragility*

TABLE 1—REGIONAL DIFFERENCES,
1970–1995: VOLATILITY

A. Currency Crises

Indicators	Volatility		
	Latin America	East Asia	Others
M2 multiplier	28.1	8.0*	11.1*
Domestic credit/GDP	21.4	12.6*	5.6*
Real interest rate on deposits	4.3	0.9*	0.8*
Ratio of lending-to-deposit interest rate	16.6	8.9*	9.7*
"Excess" real M1 balances	3.3	0.4*	0.6*
Bank deposits	321.4	6.9*	20.1*
M2/foreign-exchange reserves	76.4	35.0*	22.3*
Exports	28.4	24.9	17.4*
Imports	35.9	27.4	21.9*
Real exchange rate	36.1	38.2	7.8*
Terms of trade	19.6	14.3	4.7*
Foreign-exchange reserves	71.7	42.5*	33.6*
Domestic–foreign real-interest-rate differential	4.3	0.9*	0.9*
Output	10.4	8.9	6.1*
Stock prices	64.7	36.1*	30.6*

B. Banking Crises

Indicators	Volatility		
	Latin America	East Asia	Others
M2 multiplier	18.1	5.3*	9.5
Domestic credit/GDP	13.3	5.7*	7.0
Real interest rate on deposits	2.1	0.7*	0.7*
Ratio of lending-to-deposit interest rate	13.9	10.4	10.1
"Excess" real M1 balances	3.0	0.3*	0.4*
Bank deposits	350.2	7.3	20.8
M2/foreign-exchange reserves	61.5	15.8*	24.0
Exports	23.1	19.2	15.6*
Imports	25.1	16.4*	16.9*
Real exchange rate	32.8	26.1	11.3*
Terms of trade	16.4	12.4	4.6*
Foreign-exchange reserves	53.6	19.2*	31.1*
Domestic–foreign real-interest-rate differential	2.2	0.7*	0.8*
Output	5.7	14.0	5.0
Stock prices	80.1	34.3*	60.3

Note: Volatility is measured by the mean absolute deviation from tranquil periods (as a percentage) for each indicator in the 18 months prior to the crisis.

* Statistically significant at the 5-percent level.

TABLE 2—REGIONAL DIFFERENCES,
1970–1995: FRAGILITY

A. Percentage of currency crises			
Percentage “anomalous” ^a	Latin America	East Asia	Others
80–100	45	29	32
60–79	39	42	39
40–59	8	29	18
20–39	8	0	11
<20	0	0	0

B. Percentage of banking crises			
Percentage “anomalous”	Latin America	East Asia	Others
80–100	46	25	38
60–79	46	75	25
40–59	8	0	25
20–39	0	0	12
<20	0	0	0

^a Percentage of the indicators showing “anomalous” behavior preceding the crisis.

of an economy on the eve of a crisis, our basic premise is that the more widespread the economic problems are, the larger should be the number of indicators that exhibit “anomalous” behavior on the eve of a crisis.⁴ Thus, we need to tally, crisis-by-crisis, what proportion of the 15 indicators were showing aberrant behavior in the 24 months preceding the financial crisis. This information is summarized for the three regions in Table 2. For instance, in the 24-month period prior to Venezuela’s currency crisis in May 1994, 10 of the 15 (67 percent) of the indicators were exhibiting “anomalous” behavior; this crisis gets counted in the second row of Table 2, labeled 60–79 percent. Hence, the top row of each panel in Table 2 provides information on the share of the crises in our sample that were preceded by abnormal behavior in at least 80 percent of the indicators. Quite clearly, LA economies were more frail on the eve of crises than were economies in other regions. In 45

TABLE 3—REGIONAL DIFFERENCES, THEN AND NOW:
SEVERITY INDEX

A. Currency Crises			
Period	Severity index		
	Latin America	East Asia	Others
1970–1994	48.1	14.0	9.0
1995–1997	25.4	40.0	NA

B. Banking Crises			
Period	Severity index		
	Latin America	East Asia	Others
1970–1994	21.6	2.8	7.3
1995–1997	8.3	15.0	NA

Notes: See text for a description of the severity index. As to the severity of the banking crises for the 1997 crises in Asia, we rely on estimates of the bailout costs as of December 1997. NA denotes not applicable.

percent of the currency crises in LA, 80–100 percent of the indicators were indicating problems. Only 29 percent of the EA crises were preceded by so many flashing red lights. A similar regional disparity is evident on the eve of past banking crises.

To measure the *severity* of a currency crisis, we focus on a composite measure that averages reserve losses and the real exchange-rate depreciation. For reserves, we use the six-month percentage change prior to the crisis month, as reserve losses typically occur prior to the devaluation (if the attack is successful). For the real exchange rate, we use the six-month percentage change following the crisis month, because large depreciations occur after, and only if, the central bank concedes by devaluing or floating the currency. This measure of severity is constructed for each currency crisis in our sample, and the regional averages are reported in Table 3. For banking crises, we use the bailout costs, as a percentage of GDP, as the measure of severity; regional averages are also reported.

The first row of each panel in Table 3 presents evidence of the historical patterns; it is clear that the LA financial crises were far more severe than those elsewhere. The average severity index for currency crises is more than

⁴ For a detailed description of the methodology used to classify what is considered “anomalous” behavior in an indicator and what is not, see Kaminsky and Reinhart (1996).

three times larger for LA than for EA; an even larger discrepancy is evident for the banking crises, where the average cost of the bailout is about seven times larger in LA than in EA. The second row in each panel records the readings of this index for the recent crises, Argentina and Mexico in 1994–1995 and the ongoing crises in Indonesia, Malaysia, Philippines, and Thailand. The bailout costs of the banking sector are estimated to range from a low of about 7 percent of GDP for the Philippines to over 20 percent of GDP for Thailand. The picture that emerges is quite distinct from the historical pattern. Both on the currency and banking side, the severity of the recent crises is at par with those recorded for LA in the past. For instance, for the Thai case, by mid-December the baht had depreciated about 80 percent while the percentage decline in reserves was of a similar magnitude, when the forward position is included. The cost of the bailout of banks has surpassed that of rescuing the Mexican banks (see Amar Bhattacharya et al., 1997).

II. Why Are Regional Differences Eroding?

We have argued that, historically, financial crises have been more frequent and severe in LA than in other regions. Yet we conclude that the severity of the recent crises in EA matches LA standards more closely than the EA historical norm. In this section, we speculate why the regional differences may be eroding.

In the early 1990's, when capital began to flow to emerging market economies, countries in East Asia were enjoying substantial amounts of foreign direct investment (FDI) and a low share of their capital inflows were short-term. By contrast, Latin America's poor track record of chronic inflation and low growth was cited as a key reason why a dominant share of the capital flowing to that region was of a short-term nature and FDI flows were comparatively scarce (Table 4). By 1996, these stark regional differences in the composition of capital flows had all but disappeared. In several East Asian countries, persistent bouts of sterilized intervention kept short-term interest rates high relative to international levels and acted as a magnet for short-term capital inflows. Indeed, the evidence

TABLE 4—EAST ASIA AND LATIN AMERICA:
SIGNS OF CONVERGENCE?

Indicator	1986–1995		1996	
	East Asia	Latin America	East Asia	Latin America
Inflation rate	6.3	429.2	5.6	10.9
Real GDP growth	7.0	3.3	7.1	4.9
FDI/(short-term and portfolio flows) ^a	77.2	42.2	61.8	110.5

Note: Latin America here comprises the four largest capital importers in the 1990's: Argentina, Brazil, Chile, and Mexico.

^a The ratio of FDI to short-term and portfolio flow, expressed as a percentage. The composition of capital flows is for 1990–1992 or the “early wave” of the capital inflow surge.

presented in Peter J. Montiel and Reinhart (1998) suggests that these policies played a key role in explaining the rising volume of short-term flows that countries such as Malaysia and Thailand attracted. At the same time, a number of Latin American countries implemented major inflation-stabilization programs, and growth in that region rebounded from its bleak performance during the 1980's. Despite the setbacks associated with the Mexican crisis of December 1994 and its “tequila” effects, the pattern of lower inflation and higher growth has persisted in 1996 and 1997. Largely owing to improved economic prospects in LA, FDI began to account for a rising share of capital flows to the region.

In EA, the rising volume of short-term flows was largely intermediated by the poorly regulated and ill-supervised domestic banking sectors. Indeed, the overlending and asset-price cycles in EA of the 1990's are reminiscent of the cycles that followed financial liberalization in many Latin American countries. The policy dilemma that has recently characterized several Asian countries, the choice between high interest rates to defend the peg and ample liquidity to help troubled financial institutions, is also reminiscent of Chile's currency and banking crises in the 1980's. It appears that the combination of volatile international capital and weaknesses in

the financial sector may be at the heart of the 1997 crises in EA and may explain their severity.

III. Concluding Remarks

Historically, there were marked differences between the fierceness of financial crises in EA and in LA. In the 1990's, LA made progress toward stabilization—although at the time of this writing LA remains vulnerable to contagion. At the same time, the severity of the 1997 EA financial crises has escalated to magnitudes not seen earlier in that region and is comparable to that of the LA financial crises of the past. Regional differences are eroding.

What accounts for convergence is food for future research. However, on the basis of our analysis, one may speculate that some of the past differences in the volatility of the capital account and financial sector have diminished in the world of mobile capital and more deregulated financial markets. Regional differences in the composition of capital flows to the two regions eroded throughout the 1990's, as an increasing volume of short-term capital was funneled into Asia. Also, the booms in lending and asset prices that characterized the EA economies before the bubble burst in 1997 are reminiscent of the post-financial-liberalization episodes in LA. It appears that, in a deregulated

world, the "well-behaved" Asian financial crises are a relic of the past.

REFERENCES

- Bhattacharya, Amar; Claessens, Stijn and Hernandez, Leonardo. "Recent Financial Market Turbulence in Southeast Asia." Mimeo, The World Bank, Washington, DC, October 1997.
- Edison, Hali J. and Miller, Marcus. "The Hong Kong Handover: Hidden Pitfalls?" Mimeo, Board of Governors of the Federal Reserve, Washington, DC, July 1997.
- Kaminsky, Graciela L., and Reinhart, Carmen M. "The Twin Crises: The Causes of Banking and Balance of Payments Problems." Mimeo, Board of Governors of the Federal Reserve, Washington, DC, August 1996.
- McKinnon, Ronald and Pill, Huw. "Credible Liberalizations and International Capital Flows: The Overborrowing Syndrome." Mimeo, Stanford University, 1994.
- Montiel, Peter J. and Reinhart, Carmen M. "The Dynamics of Capital Movements to Emerging Economies During the 1990s," in S. Griffith-Jones and M. Montes, eds., *Short-term capital movements and balance of payments crises*. Oxford: Oxford University Press, 1998 (forthcoming).

On the Importance of the Precautionary Saving Motive

By ANNAMARIA LUSARDI*

The life-cycle-permanent-income model has been the primary theoretical framework for research on saving. The basic intuition of the model is that households should smooth consumption over the life cycle. They should, therefore, save prior to retirement to offset the decline in future income and start drawing down wealth when they retire. A promising extension to the life-cycle model, which has become known as the theory of precautionary saving, has emphasized that saving serves not only to spread resources over the life cycle, but also to insure against uncertain events, such as shocks to income. This theory provides many useful insights about the behavior of saving and may reconcile some of the puzzles that confront the saving literature.

While the theory is promising from a theoretical point of view, the empirical work faces many difficulties. The central difficulty is how to obtain a good measure of risk. As discussed at length in Martin Browning and Lusardi (1996), one needs to identify some observable and exogenous sources of risk that vary significantly across the population. Some authors, such as Christopher Carroll and Andrew Samwick (1995), use the variance of income from observed income processes to proxy for risk. However, this approach is sensitive to the presence of measurement error in income and how much the consumer knows that the econometrician does not. Other authors, such as Jonathan Skinner (1988), have used other proxies for risk such as the occupation of the head of the household. This can be unsatisfactory if people select themselves into occupations on the basis of their degree of risk

aversion. A new and innovative approach has been used by Luigi Guiso et al. (1992). They use data on the subjective probability distribution of future income from a sample of Italian households. One of the problems with their work, however, is that these types of questions might not be easy for respondents to understand.

In this paper, I use data from a new data set: the Health and Retirement Study (HRS), that provides data on subjective probabilities of job loss. The question in the HRS is rather simple and intuitive. As far as I know, this paper is the first application of subjective data in the estimation of a precautionary-saving model in the United States. An additional contribution is that, by looking at a group of the population (people close to retirement) for whom income risk should be relatively small, I can assess the importance of the precautionary saving motive.

I. A Model of Precautionary Saving

One of the principal lessons learned in the past decade is that predictions from a life-cycle-permanent-income model are less general than previously thought and can change quite dramatically when uncertainty is taken into account. Once it is assumed that income is not certain and can vary substantially over the working life, and that consumers dislike uncertainty, saving behavior becomes richer than in traditional models.¹

The theoretical predictions can be summarized with reference to the following reduced-form equation, which has been estimated by many authors:

$$(1) \quad \frac{W_h}{Y_h^p} = f(\text{age}, \mathbf{X}_h, \sigma_h^2).$$

* Department of Economics, Dartmouth College, Rockefeller Hall, Hanover, NH 03755. I thank Rob Alessie, Patty Anderson, Gary Engelhardt, Nancy Jianakoplos, and participants at the junior lunch at Dartmouth College for suggestions and comments. Financial support from the Olin Foundation and the National Institute on Aging (grant no. 1-R01-AG13893-01A1) is gratefully acknowledged.

¹ See the review of precautionary saving in Browning and Lusardi (1996) and Angus Deaton (1992).

In the precautionary-saving model, wealth divided by permanent income (W/Y^p) of household h is a function of age, and other household characteristics (X), that reflect preferences parameters and may include permanent income if preferences are nonhomothetic. One additional term that affects wealth accumulation is the uncertainty about income, as measured by the variance (σ^2). Note that there is a positive association between uncertainty and wealth; the higher the uncertainty, the higher is the accumulation. Even though it is very stylized, the expression above brings out the main intuition of the model. Precautionary saving can be seen as an extension of the basic intertemporal optimizing framework. Wealth continues to depend on permanent income as in traditional models of saving, but there are additional variables, such as the variance of income, that also play a role in explaining accumulation.

II. Empirical Analysis of the Precautionary-Saving Model

In the empirical work, I use data from the first wave of the HRS, a new survey conducted at the University of Michigan which was started in 1992. As examined in some detail below, this survey provides detailed information on wealth and the retirement process, with a focus on health, labor-market, and economic and psychosocial factors. The age range of respondents is restricted to 51–61 years old. In addition, the HRS oversamples minorities and residents in Florida. It is only the individual deemed most knowledgeable about the family's assets, debts, and retirement planning, who is asked questions on housing, net worth, and income of the family (respondent hereafter). A thorough examination of the quality of the HRS data and a comparison with other data sets is reported in Thomas Juster and James Smith (1994) and Smith (1995).

As illustrated in equation (1), to estimate a precautionary-saving model one needs information on three factors: wealth, household characteristics (including permanent income), and income risk. I comment hereafter on each factor and the information reported in the HRS.

(i) *Wealth*.—Two measures of wealth are used in the empirical work. The first measure

(financial net worth) is defined as the sum of checking and saving accounts, bonds, stocks, IRA's, and other assets, minus short-term debt. The second measure (total net worth) is obtained by adding financial net worth to home equity, other real estate, business equity, and vehicles. Even though one expects precautionary accumulation to be mainly in liquid assets, for some assets, such as IRA's, there are provisions for withdrawals due to emergencies. Additionally, the development of the home-equity line of credit in the late 1980's and the fact that assets, such as vehicles and housing, can be used as collateral makes it worth considering a comprehensive measure of household resources such as total net worth.

To examine wealth more closely, in Table 1 I consider the sample of households in the first wave of the HRS and report weighted statistics of financial and total net worth (I exclude some major outliers). Both measures have a wide distribution. Note that the mean is well above the median and that the distributions of financial and total net worth are highly skewed to the right. Furthermore, many households arrive close to retirement with a small amount of assets, both in levels and as a ratio of their current income.

(ii) *Household characteristics and preference parameters*.—The richness of information in the HRS allows a careful investigation into the extent of household wealth accumulation. The HRS provides data on many household characteristics that are useful in the estimation of the precautionary-saving model (sex, gender, education, and health status). An innovation with respect to other surveys is the provision of data on a set of individual characteristics, from which one can gain information on preferences. For example, the HRS questionnaire contains some unique questions that allow the researcher to evaluate the attitude toward risk and calculate a measure of the coefficient of risk aversion.² It also reports information on the planning horizon in making financial decisions,

² See the evaluation of the risk aversion measure in the HRS in Robert Barsky et al. (1997).

TABLE 1—DISTRIBUTION OF FINANCIAL AND TOTAL NET WORTH

Percentile	Financial wealth	Total net worth	Net worth/income
5	-4,300	0	0
10	-1,000	1,500	0.08
25	100	32,000	0.98
50	14,950	102,000	2.47
75	66,000	237,100	5.30
90	173,000	484,000	10.94
95	286,000	752,000	16.84
Mean:	63,441	206,615	4.60
(SD):	(136,505)	(331,774)	(6.84)

Source: Author's calculations from the HRS.

which could be considered an index of time preferences.³ In addition, respondents are asked to assess their chances of surviving up to the age of 75 or 85, whether they expect their earnings to go up or down in the future, and whether they expect to leave a bequest. These variables are important in accounting for the wide variation in wealth that we observe in the data.

(iii) *Income risk.*—An additional innovation in the HRS is the provision of subjective data that can be used to calculate a measure of income risk. Individual respondents are asked to evaluate the chances that they will lose their jobs in the next year. The question is as follows: "Sometimes people are permanently laid off from jobs that they want to keep. On a scale from 0 to 10 where 0 equals absolutely no chance and 10 equals absolutely certain, how likely is it that you will lose your job during the next year?"⁴

The responses to this question are sensible.⁵ Many respondents report that the odds of losing their jobs next year are low, with

TABLE 2—PROBABILITY OF JOB LOSS IN UPCOMING YEAR

Variable	Estimates	Standard error
Age of respondent	-0.004	0.005
Union membership	-0.093	0.046
Past unemployment	0.292	0.042
Less than 30-week work year	0.382	0.145
Less than 25-hour work week	0.143	0.080
Tenure at current employer	-0.009	0.002

Note: This table reports the estimates from an ordered probit regression of the subjective probabilities of job loss on the variables listed in the first column.

the majority of respondents choosing values between 0 and 0.5. In Table 2, I report the results of an ordered probit regression of the probabilities of job loss on a set of job characteristics.⁶ The results are as expected. For example, the odds of job loss are positively related to past unemployment, and negatively related to union membership and the number of years spent at the respondent's current employer.

By making some simple assumptions, it is possible to use this information to derive a measure of income variance that can be used in the estimation of the precautionary-saving model.⁷ If the unemployment-insurance replacement rate is zero, and there are no changes in earnings if the respondent does not lose his/her job, then it is easy to show that the variance of earnings is equal to $p(1-p)Y^2$, where p is the subjective probability of losing the job and Y is earnings. If the replacement rate is equal to α , the variance of income becomes $p(1-p)(1-\alpha)^2Y^2$.

³ The planning horizon is listed as follows: next few months, next year, next few years, next 5–10 years, longer than 10 years.

⁴ After rescaling to 0–1, the responses can be interpreted as subjective probability distributions of the event.

⁵ Other authors have examined subjective data in the HRS. See the evaluation of the subjective probability of survival in Michael Hurd and Kathleen McGarry (1995).

⁶ These estimates refer to the final sample chosen for the empirical estimation of the precautionary saving motive, but results are similar when taking the total sample of respondents in the HRS.

⁷ Income variance in this case refers to earnings variance.

TABLE 3—PRECAUTIONARY ACCUMULATION:
TOTAL NET WORTH

Variable	OLS estimates (SE)	Median estimates (SE)
Constant	-7.578 (4.379)	-3.832 (5.567)
Age	0.241 (0.163)	0.065 (0.207)
Age-squared/100	-0.174 (0.152)	-0.005 (0.193)
Male	-0.441 (0.074)	-0.305 (0.094)
White	0.447 (0.115)	0.382 (0.147)
Black	-0.418 (0.126)	-0.242 (0.161)
High school	0.251 (0.102)	0.307 (0.129)
Some college	0.271 (0.135)	0.231 (0.172)
College	0.478 (0.182)	0.340 (0.231)
More than college	0.683 (0.264)	0.405 (0.335)
Married	0.400 (0.182)	0.558 (0.232)
Living with partner	-0.021 (0.263)	0.181 (0.335)
Divorced	-0.091 (0.164)	0.046 (0.209)
Widowed	0.519 (0.185)	0.616 (0.236)
Separated	-0.055 (0.222)	0.050 (0.282)
Excellent health	1.040 (0.231)	0.674 (0.290)
Very good health	0.684 (0.227)	0.516 (0.285)
Good health	0.608 (0.225)	0.495 (0.283)
Fair health	0.486 (0.235)	0.325 (0.296)
High risk aversion	0.225 (0.101)	0.408 (0.128)
Moderate risk aversion	0.385 (0.129)	0.492 (0.165)
Low risk aversion	0.242 (0.133)	0.331 (0.170)
Short planning horizon	0.224 (0.083)	0.149 (0.106)
Medium planning horizon	0.199 (0.086)	0.279 (0.110)
Long planning horizon	0.426 (0.132)	0.492 (0.168)
Probability of living to age 75	-0.008 (0.120)	0.007 (0.153)
Bequest motive	0.985 (0.065)	0.830 (0.084)
Income expected to fall	0.341 (0.127)	0.205 (0.161)
Income expected to rise	-0.085 (0.070)	-0.055 (0.088)
Past unemployment	-0.297 (0.069)	-0.274 (0.089)
Permanent income/ 1,000	-0.003 (0.004)	0.0004 (0.005)
Variance of income	0.009 (0.005)	0.038 (0.006)

Notes: This table reports ordinary least-squares (OLS) and median estimates from the regression of the ratio of total net worth to permanent income on the variables listed in the first column. Standard errors are in parentheses. The R^2 values of the OLS and median regressions are 0.165 and 0.104, respectively. The number of observations is 3,391.

III. Empirical Estimates

The empirical estimates of a regression of wealth divided by permanent income on a set of household characteristics, permanent income, and the variance of income are reported in Table 3.⁸ To construct the final sample, I

⁸ I construct a measure of permanent income by regressing household income on a set of demographic and

delete those households whose head is retired (fully or partially), or planning to retire in the current year, since earnings risk is not relevant for these respondents. The sample is also much reduced by the fact that many respondents are not asked the question about job loss.⁹ Since wealth has such a wide distribution and outliers can affect the estimates, I trim the distribution and exclude the top and bottom 2 percent of the distribution. Additionally, I use estimators such as median estimators that can take better account of the outliers.

The empirical results concerning household characteristics and preferences are consistent with the basic theory of intertemporal optimization. For example, households whose respondent is more risk-averse accumulate more wealth. Respondents with long planning horizons (lower rates of time preference) accumulate more. In accordance with the prediction that saving occurs to offset future declines in income, households that expect their earnings to go down in the future accumulate more wealth. In addition, households whose head suffered periods of unemployment in the past have significantly lower wealth. It is important to account for households that have already suffered unemployment shocks in the past. These households are likely to have lower wealth (shocks depleted their stocks of assets), but also a higher risk of losing jobs in the future.

In accordance with the theory of precautionary saving, the sign of the variance of income is positive and statistically significant, indicating that people who face a higher income risk

firm characteristics. I use age, sex, and marital status in addition to education and occupation dummies which are interacted with age. I also use dummies for whether the respondent works in a small firm (fewer than 20 employees), for whether the respondent belongs to a union, and for whether the respondent works full time. Since the age range is only 10 years in the HRS, I have not accounted for cohort effects in income. The predictions from this regression are used as a proxy of the permanent component of income. Note also that I normalize the variance of income by dividing by permanent income.

⁹ This question is only asked to respondents who "work for someone else" at the time of the interview. Note that I also delete respondents with missing data on the variables of interest, and respondents outside the 51-61 age range.

save more and accumulate more wealth. The estimates are statistically significant across estimation methods and increase in magnitude when considering median regressions, which are less affected by outliers. However, the contribution of precautionary saving to wealth accumulation is not very large. Evaluated at the sample means, the coefficient estimates show that the extent of precautionary accumulation (measured by the ratio of total net worth to permanent income) ranges from 1 percent to 3.5 percent. I also have performed the estimation on financial net worth. In that case, the extent of precautionary accumulation ranges from 2 percent to 4.5 percent.¹⁰

Even though I have assumed that the replacement rate is zero, I have also experimented with values of the replacement rate that range from 0.50 to 0.70, and results do not change substantially. The estimates can, however, be made more precise. It is clear that households with only one earner are much more exposed to risk than households where both spouses work. To account for this fact, I have interacted the variance of income with a dummy equal to 1 if there is more than one earner in the household and added that dummy separately in the regression. The predictions of the theory are verified in the data. Both the coefficient of the variance of income and that of interaction term are statistically significant, and the sign of the interaction term is, as expected, negative. The ordinary least-squares estimates of the variance of income and the interaction term are 0.045 (SE = 0.011) and -0.043 (SE = 0.012), respectively.

IV. Concluding Remarks and Further Research

Empirical estimates using HRS data indicate that the variance of income has a role in explaining saving and wealth accumulation of people close to retirement. Consistent with other findings in the literature, precautionary saving does not provide a rationale for a large accumulation of wealth and certainly cannot explain the wealth holdings of the very rich. However, the fact that

there is evidence in favor of precautionary saving in this group of the population provides a strong indication that this motive is important. Further research in this area will consider other sources of risk that can affect saving. Apart from income risk, health and longevity risk can also be important and can provide useful insights to explain the wealth holdings of many households in the United States.

REFERENCES

- Barsky, Robert B.; Juster, Thomas F.; Kimball, Miles S. and Shapiro, Matthew D. "Preference Parameters and Behavioral Heterogeneity: An Experimental Approach in the Health and Retirement Study." *Quarterly Journal of Economics*, May 1997, 62(2), pp. 537-79.
- Browning, Martin and Lusardi, Annamaria. "Household Saving: Micro Theories and Micro Facts." *Journal of Economic Literature*, December 1996, 34(4), pp. 1797-1855.
- Carroll, Christopher D. and Samwick, Andrew A. "How Important is Precautionary Saving?" National Bureau of Economic Research (Cambridge, MA) Working Paper No. 5194, July 1995.
- Deaton, Angus. *Understanding consumption*. Oxford: Oxford University Press, 1992.
- Guiso, Luigi; Jappelli, Tullio and Terlizzese, Daniele. "Earnings Uncertainty and Precautionary Saving." *Journal of Monetary Economics*, November 1992, 30(2), pp. 307-38.
- Hurd, Michael D. and McGarry, Kathleen. "Evaluation of the Subjective Probabilities of Survival in the Health and Retirement Study." *Journal of Human Resources*, Supplement 1995, 30, pp. S268-92.
- Juster, Thomas F. and Smith, James P. "Improving the Quality of Economic Data: Lessons from the HRS." Mimeo, Institute for Social Research, University of Michigan, 1994.
- Skinner, Jonathan. "Risky Income, Life Cycle Consumption, and Precautionary Savings." *Journal of Monetary Economics*, September 1988, 22(2), pp. 237-55.
- Smith, James P. "Racial and Ethnic Differences in the Health and Retirement Study." *Journal of Human Resources*, Supplement 1995, 30, pp. S159-93.

¹⁰ For brevity, estimates are not reported but are available from the author upon request.

Risk, Entrepreneurship, and Human-Capital Accumulation

By MURAT F. IYIGUN AND ANN L. OWEN*

An economy's human-capital stock is determined by both entrepreneurs and professionals. Entrepreneurs provide the economy with new ideas, products, and ways of doing things, while professionals utilize their accumulated knowledge to facilitate economic transactions. Both skills are necessary for a healthy economy. Yet, while professional and entrepreneurial skills can complement each other in aggregate production, they can compete for an individual's time in their accumulation.

In addition, while both entrepreneurial and professional skills can arguably be defined as "human capital" there are important dimensions along which entrepreneurship and professionalism differ. We find that considering these differences leads to important implications for the dynamic behavior of an economy's human-capital stock. In particular, we show that entrepreneurial human capital plays a relatively more important role in intermediate-income countries, whereas professional human capital is relatively more abundant in higher-income economies. We also show that, in an economy where both entrepreneurial and professional human capital affect the future level of technology, the initial stocks of both types of human capital are important for development. Finally, we demonstrate an important implication of considering more than one means of accumulating human capital: an inefficient allocation of time between schooling and gaining entrepreneurial experience may occur.

Our model is premised on one fundamental idea: as an economy develops, improvements in the aggregate production technology

broaden an individual's opportunities within the economy. These changing opportunities alter the relative costs of individual choices, thus affecting the outcome of individual decision-making. More specifically, in our model, individuals choose to allocate fewer resources toward entrepreneurship in a more developed economy because good, safe alternatives to this risky activity exist.

We are motivated in part by the observation that, as an economy develops, its occupational structure changes. More people become employed by others. Figure 1 shows that higher levels of per capita income are associated with a lower ratio of employers to employees. The model we describe below generates such a result, with more people choosing the relatively safe return of schooling over relatively risky self-employment as per capita income grows.

Abhijit Banerjee and Andrew Newman (1993) also explore this pattern of occupational change and development. They focus, however, on the importance of an individual's financial resources and access to credit in the decision to become an entrepreneur. We take a slightly different view on the defining characteristic of entrepreneurship, choosing to focus on the elements of risk and relative return inherent in the concept. Thus, while Banerjee and Newman demonstrate that economic development may be associated with increased entrepreneurship, our model shows that, in more developed economies, the increase in the return to professional activities relative to that of inherently risky entrepreneurial ventures has an offsetting negative effect. While both approaches generate more entrepreneurs *in total* in a developed economy, our approach shows that economic development will be associated with a decline in the number of entrepreneurs *relative* to professionals.

I. Key Features of the Model

There are a few key features of our model that are useful to highlight at the outset. Most

* Board of Governors of the Federal Reserve System, Washington, DC 20551, and Hamilton College, 198 College Hill Road, Clinton, NY 13323, respectively. This paper represents the views of the authors and should not be interpreted as reflecting those of the Board of Governors of the Federal Reserve System or other members of its staff.

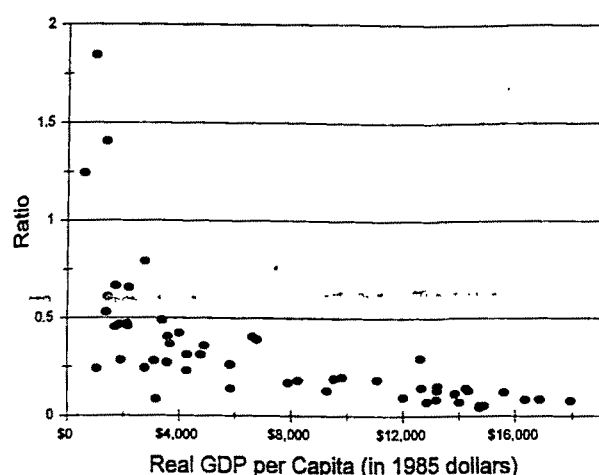


FIGURE 1. REAL GDP PER CAPITA AND THE EMPLOYERS-TO-EMPLOYEES RATIO

Notes: The figure excludes agricultural workers. Employers include those who are self-employed. One observation per country is reported within the period 1986–1992.

Sources: Penn World Tables (Mark 5.6) (Robert Summers and Alan Heston, 1991) and the 1993 *Year Book of Labour Statistics*.

important, we consider two types of human capital: professional and entrepreneurial. These two types of human capital differ in three important ways. First, entrepreneurship is risky. There is a positive probability that entrepreneurial activity will result in failure, an outcome with a very low payoff. In contrast, professional activity provides a safe return. Second, entrepreneurial and professional skills are augmented in different ways. Entrepreneurs learn by doing; entrepreneurial skills are honed by investing time working in an entrepreneurial venture. Professionals, on the other hand, accumulate their skills primarily by investing time in schooling. Thus, increased investment in professional skills competes for an individual's time and "crowds out" entrepreneurial investment. Finally, we assume that entrepreneurial and professional skills are not perfect substitutes in producing the aggregate production technology. Both skills are necessary, but one is potentially more important in determining the level of technology available in the economy.

Another key feature of our model is that the production technology employed in one time period is a function of the previous generation's human capital (entrepreneurial and pro-

fessional). This element contributes to the dynamics of the model, providing a link between today's human-capital stock and tomorrow's.

The last important feature of our model is really the absence of a common assumption. Models that treat all types of human capital as one implicitly make the assumption that different types of human capital are combined in the same way in output production as they are combined in the formation or adoption of technology. We do not make this assumption. Instead, we allow the relative importance of each type of human capital to vary by activity. In our model, entrepreneurial human capital may be more productive in the development of technology, and professional human capital may be more productive in the utilization of the existing technology, or vice versa. This, combined with the assumption above, creates a classic externality in which individuals in one generation will not select the socially optimal level of schooling and work experience. In contrast to the usual conclusion, however, in our setup, individuals may choose too much education and not enough work experience if entrepreneurship is an important determinant of the level of technology.¹

II. Overview of the Model

We employ a two-period overlapping-generations model. There is no population growth. In the first period of life, individuals of differing abilities decide whether to become human-capital providers or raw-labor providers. Ability increases the wages paid to both professionals and entrepreneurs for a given level of technology.² If individuals decide to become human-capital providers, they must allocate their time between going to school to accumulate professional human capital and

¹ Joseph A. Schumpeter (1934) argues that entrepreneurs completely determine the level of technology. This is a special case of our model in which too much education will always be chosen.

² In our model, human-capital acquisition is related to ability. It would be straightforward to adapt our model to other commonly used acquisition rules, such as the one found in Owen and David N. Weil (1998) which takes into account parental wealth and ability.

working in an entrepreneurial venture to accumulate entrepreneurial skills. There are diminishing returns to time spent in school or at work, ensuring that individuals always accumulate some of both skills. Individuals who choose not to become human-capital providers work as raw laborers during this initial period. In the second period of life, individuals work and consume. Their utility depends on their second-period consumption.

Aggregate production is governed by a constant-returns-to-scale production function that features skill-biased technological change. The level of technology is a function of both the entrepreneurial and professional human capital of the previous generation. As the economy develops and technology evolves, wages to both professionals and entrepreneurs increase relative to the wages of raw labor.³ This increased incentive to become a human-capital provider is an important element of the dynamics of the model. It ensures a growing pool of human-capital providers who develop their entrepreneurial and professional skills.

When making the choice between accumulating professional and entrepreneurial skills, individuals must consider the fact that the payoff to entrepreneurship is risky. There is a positive probability that entrepreneurs will fail and receive a zero return on their entrepreneurial skills. The payoff to successful entrepreneurs, however, does increase as an economy develops, and the expected value of entrepreneurship increases with per capita income.

However, as technology improves, the return to being a professional increases as well. As the return to this safe activity increases, human-capital accumulators devote an increasing proportion of their time to accumulating professional skills in spite of the fact that the expected value of entrepreneurship has increased. Essentially, as professional opportunities increase, individuals in more developed economies face a higher opportunity

cost of becoming an entrepreneur. And although individuals contemplating becoming an entrepreneur in a more developed economy face a clearly better lottery than those in a less developed economy, the price of the lottery ticket—forgone professional earnings—is also higher, making fewer individuals willing to take the bet.

III. Results

This model has some important implications for the dynamic behavior of the human-capital stock. One of the main results is that entrepreneurial human capital is relatively more important in intermediate-income countries, whereas professional human capital is relatively more abundant in richer economies. Thus, as an economy grows, individuals who provide human capital choose to allocate an increasing proportion of their time to the accumulation of professional skills. Yet, even though the relative stocks of professional and entrepreneurial skills change, the absolute magnitudes of both skills increase with per capita income. The skill-biased technological change that accompanies growth creates increasing incentives for individuals to become human-capital providers and lowers the ability level that makes human-capital investment profitable. Thus, the total stock of entrepreneurial skills increases with per capita income as the pool of human-capital providers grows larger. The relative stocks of entrepreneurial and professional skills in the steady state will depend on the marginal productivity of each skill in aggregate production as well as the probability of entrepreneurial success.

Another result of our model is that initial stocks of both types of human capital are important for the process of development. Economies that have too little of either entrepreneurial or professional human capital may end up in a development trap with very low investment of human capital of any type. A relevant example of the importance of this second result may be found in the former East-Bloc countries. As some have pointed out (e.g., Jody Overland and Michael Spagat, 1996), these economies have highly educated labor forces and may be positioned very well for strong economic growth. However, our

³ With this formulation, we abstract from how the magnitude of improvements in the level of technology (due to higher human-capital stock) might affect relative returns to different components of human capital. For a relevant discussion on this issue, see Oded Galor and Daniel Tsiddon (1997).

model makes a more pessimistic prediction; future growth of these economies will be constrained by their labor forces' lack of entrepreneurial experience.

Finally, our third result is that, because the social marginal returns to work and education may differ from the private marginal returns, an inefficient allocation of resources between working and schooling will result. The nature of the inefficiency will depend on exactly how entrepreneurial and professional skills contribute to the level of technology. The cause of the inefficiency may also shift over the process of development. If entrepreneurial skills are relatively more (less) important in determining technology, the steady state will have too much (not enough) education. The inefficiency, of course, results not from too much human capital, but from a misallocation of the existing human-capital stock between professional and entrepreneurial activities. In fact, a more efficient ratio of professional and entrepreneurial skills will raise the steady-state level of technology, the wages paid to human-capital providers, and therefore, the economy's human-capital stock.

IV. Conclusion

The model we describe above highlights the importance of considering more than one type of human capital for the evolution of an economy. It shows how the incentives to accumulate entrepreneurial human capital change as the opportunity cost of entrepreneurship (forgone professional earnings) increases. It ex-

plains why professional human capital takes on greater relative importance in developed economies and suggests that, when more than one type of human capital exists, individuals may not allocate their resources efficiently among the alternatives.

REFERENCES

- Banerjee, Abhijit and Newman, Andrew. "Occupational Choice and the Process of Economic Development." *Journal of Political Economy*, April 1993, 101(2), pp. 274-98.
- Galor, Oded and Tsiddon, Daniel. "Technological Progress, Mobility, and Economic Growth." *American Economic Review*, June 1997, 87(3), pp. 363-82.
- Overland, Jody and Spagat, Michael. "Human Capital and Russia's Economic Transformation." *Transition*, June 1996, 13(2), pp. 12-15.
- Owen, Ann L. and Weil, David N. "Intergenerational Earnings Mobility, Inequality, and Growth." *Journal of Monetary Economics*, February 1998, 41(1), pp. 71-104.
- Schumpeter, Joseph A. *The theory of economic development*. Cambridge, MA: Harvard University Press, 1934.
- Summers, Robert and Heston, Alan. "The Penn World Table: An Expanded Set of International Comparisons, 1950-1988." *Quarterly Journal of Economics*, May 1991, 106(2), pp. 327-68.
- Year Book of Labour Statistics*. Geneva: International Labour Office, 1993.

THE ECONOMICS OF GUN CONTROL[†]

Who Owns Guns? Criminals, Victims, and the Culture of Violence

By EDWARD L. GLAESER AND SPENCER GLENDON*

The development of the state over the past several millenia has witnessed a transition from private to public protection of property rights. Recent analysts have argued that public enforcement of property rights is critical for economic development and for the success of the former communist bloc. Private weapons, such as guns, represent a means of privately defining property rights. This paper examines guns to understand the places where private justice still dominates public protection of life and property. While the literature on the effects of gun ownership often disagrees (see Lott, 1998), it does agree that guns matter, and we are attempting to understand the demand for guns.

Among respondents to the National Opinion Research Center's (NORC) General Social Survey (GSS), 45 percent admit to having some form of gun in the house, and 22 percent admit to having a handgun in the house. More than 50 percent of gun-owners in the GSS have more than one type of gun; Philip Cook and Mark Moore (1981) report that the average gun-owner owns 4.5 guns. Gun-owners are more likely to be middle-aged, married men with teenage children and to be high-school graduates and own their own homes. Gun-owners hunt and are much less common in big cities, among minorities, or among college graduates. Gun-owners resemble neither

criminals nor victims, although people who own only pistols and no other form of gun do seem to be responding to fear of crime.

Waiting periods have little effect on gun ownership overall, but they do reduce the share of gun-owners who have been arrested. Gun ownership is linked to average gun ownership in one's peer group and tastes for violent retribution, so community norms seem to matter. Gun-owners are also suspicious of the courts, and gun ownership is highest where police are less available, so private and public enforcement of property rights appear to be substitutes.

I. Discussion

Hunters, criminals, and people who seek guns in self-protection are predicted to demand guns. Guns may be particularly important when threats of physical violence are common in everyday social conflicts. Geoffrey Canada (1995) describes the world of inner-city youth where weapons serve as a threat point in most bargaining situations. Because police are rare and hostile, disadvantaged young men do not turn to the law when there is a dispute over property rights. Instead, they rely on private justice and weapons.

Jon Nisbett and Dov Cohen (1996) suggest that private justice is also the norm in the American South, where relying on the legal system is seen as a negative signal about individual competence. These authors argue that the Southern "culture of honor" represents a logical development of extreme vengefulness among Southern herdsmen whose capital stocks could be easily stolen. One interpretation of gun ownership is that visible weaponry serves as a signal of willingness and ability to fight. Gun ownership also may simply serve as a means of protecting one's own property,

[†] *Discussants:* Isaac Ehrlich, State University of New York-Buffalo; Douglas Weil, Center To Prevent Handgun Violence; Philip J. Cook, Duke University; Gary Kleck, Florida State University.

* Department of Economics, Harvard University, Cambridge, MA 02138; Glaeser is also affiliated with the National Bureau of Economic Research. Glaeser acknowledges financial support from the National Science Foundation and the Sloan Foundation.

but individuals with guns may also serve as enforcers of generally accepted community norms, so high-status individuals may be particularly likely to own guns.

We have four tests of the view that guns are a symptom of a "culture of private justice." First, we predict that gun ownership is higher for individuals whose peers own guns. This strategic complementarity occurs because in a fight the benefit of having a gun rises if your opponent has a gun and because the likelihood of being punished for using or having a gun declines if everyone is a gun user. Second, gun ownership should decline with police availability and confidence in the legal system. Third, following Nisbett and Cohen (1996), we expect to find a connection between a general tendency toward violent retribution and gun ownership. Fourth, since handguns provide a less visible signal, we expect to see that the three effects just described are stronger for guns generally than for handguns.

II. Results

The 1972–1994 General Social Survey (GSS) provides the largest sample size and richest set of covariates of any U.S. data set with questions about gun ownership. The GSS surveys approximately 1,500 randomly selected people annually in metropolitan and rural areas across the United States. Gun-ownership questions, including type of gun, were asked of 20,907 people in various years between 1973 and 1994. Questions about other variables of interest were asked less frequently, thereby restricting our sample size in some regressions. We focus on whether individuals have a gun in the house, rather than on whether they own a gun themselves, because this first variable was available for more years and with more detail. Additional data are drawn from the U.S. Statistical Abstract and the City and County Data Book. Our waiting-period data were generously provided by John Lott.

Table 1 shows our first set of results. Table 1A compares means for gun-owners and people who do not own guns; Table 1B reports regression coefficients for ordinary least-squares regressions of gun ownership. In the first regression, the dependent variable is own-

ing any sort of gun, and the sample is complete. In the second regression, the dependent variable is only owning a handgun, and to eliminate overlap with the first regression, the sample excludes individuals who have other forms of guns (i.e., rifles or shotguns).

The first three rows show that gun-owners are less likely to be either college graduates or high school dropouts. The next row shows that blacks are slightly less likely to own guns overall, but more likely to own pistols. Women are particularly unlikely to own guns. Marriage has a strong positive effect on gun ownership. Having children who are babies is negatively associated with gun ownership, but having children who are teenagers is positively associated with owning guns (although negatively associated with owning pistols). Gun-owners are disproportionately more than 40 years old. Gun and handgun ownership are particularly high in the South; the East has the least gun ownership. Hunters disproportionately own guns, but not pistols. Individuals who live in bigger cities are less likely to own guns.

Income is positively correlated with gun ownership, perhaps because guns are expensive or because high-income individuals have more to protect or because high incomes enable gun users to avoid any adverse consequences of their gun use. We have included a dummy variable for whether the individual refuses to admit his or her income, and this variable is correlated with gun ownership, possibly because the missing-income variable may capture general mistrust. In all subsequent regressions, we will include all of the variables in Table 1 as controls. To preserve sample sizes, we did not include home-ownership in the regression, but gun-owners are more likely to own homes (see Denise DiPasquale and Glaeser, 1998), perhaps because the desire to own a gun rises with the amount of goods to protect.

Table 2 first tests the theories about victimization, criminality, and gun ownership. The first three rows refer to results from a single regression (for each type of gun ownership) that includes variables for having been victimized in the last year, having ever been arrested for something other than a traffic offense, and fearing crime in one's neighborhood. We find no connection between these variables and gun ownership generally, but people who are afraid

TABLE 1—SUMMARY STATISTICS
AND BASELINE REGRESSIONS

A. Summary Statistics:		
Independent variable	Mean (SD)	
	Non-gun-owners	Gun-owners
Education ≥ 16 years	0.21 (0.41)	0.15 (0.36)
Education 12–15 years	0.50 (0.50)	0.57 (0.50)
Education < 12 years	0.29 (0.45)	0.29 (0.45)
Black	0.18 (0.38)	0.098 (0.30)
Female	0.63 (0.48)	0.49 (0.50)
Married	0.60 (0.49)	0.79 (0.41)
Children ages 0–6	0.19 (0.39)	0.19 (0.39)
Children ages 7–12	0.20 (0.40)	0.24 (0.42)
Children ages 13–18	0.15 (0.36)	0.21 (0.41)
Age < 30	0.26 (0.44)	0.21 (0.41)
Age 30–39	0.22 (0.42)	0.22 (0.42)
Age 40–49	0.15 (0.35)	0.19 (0.39)
Age 50–59	0.12 (0.33)	0.16 (0.37)
Age 60+	0.26 (0.44)	0.23 (0.42)
South	0.29 (0.45)	0.41 (0.49)
East	0.26 (0.44)	0.013 (0.033)
Midwest	0.26 (0.44)	0.29 (0.45)
West	0.19 (0.39)	0.17 (0.38)
Log of city population	3.98 (2.25)	2.79 (1.98)
Non-MSA resident	0.21 (0.40)	0.41 (0.49)
Log income	8.85 (3.03)	9.40 (2.71)
Missing income	0.96 (0.29)	0.07 (0.256)

B. OLS Regressions:

Independent variable	Regression coefficient (SE)	
	Own gun	Own pistol
Education ≥ 16 years	-0.103 (0.009)	-0.042 (0.008)
Education < 12 years	-0.019 (0.009)	-0.002 (0.008)
Black	-0.009 (0.01)	0.046 (0.008)
Female	-0.081 (0.007)	-0.035 (0.007)
Married	0.088 (0.009)	0.024 (0.008)
Children ages 0–6	-0.041 (0.010)	-0.024 (0.009)

TABLE 1—Continued.

B. OLS Regressions (continued):

Independent variable	Regression coefficient (SE)	
	Own gun	Own pistol
Children ages 7–12	-0.001 (0.01)	-0.000 (0.009)
Children ages 13–18	0.024 (0.010)	-0.021 (0.010)
Age < 30	-0.084 (0.012)	-0.043 (0.011)
Age 30–39	-0.054 (0.011)	-0.014 (0.011)
Age 50–59	0.015 (0.013)	0.009 (0.013)
Age 60+	0.000 (0.013)	0.028 (0.012)
South	0.081 (0.009)	0.085 (0.008)
East	-0.095 (0.010)	-0.047 (0.009)
West	0.034 (0.010)	0.030 (0.010)
Log of city population	-0.025 (0.002)	-0.004 (0.002)
Non-MSA resident	0.63 (0.009)	0.009 (0.009)
Log income	0.072 (0.004)	0.027 (0.004)
Missing income	0.680 (0.045)	0.230 (0.004)

Notes: Table 1B reports coefficients from ordinary least-squares (OLS) regressions with a dummy indicating gun ownership as the dependent variable. The own-pistol regression excludes rifle and shotgun owners. The own-gun regressions has 19,035 observations and an R^2 of 19.44 percent. The own-pistol regression has 9,556 observations and an R^2 of 16.5 percent.

buy pistols. In the next regression, we formed “predicted victimization” using the regression coefficients from regressing being victimized on demographic characteristics. We find a strong negative relationship between predicted victimization and gun ownership. The fifth row examines “predicted arrest” (similarly defined) and finds that gun-owners do not resemble arrestees either.

The next rows test the idea of a culture of violence. We constructed an average gun-ownership measure in one’s education group (college graduates, high-school graduates, or high-school dropouts) in one’s state from the GSS (excluding oneself). The first “mean ownership” row shows that there is a connection between individual and peer ownership for guns generally and for owning just a pistol. The next row shows that when we include state

TABLE 2—GROUP BEHAVIOR AND ATTITUDES

Independent variable	Regression coefficient (SE)	
	Own gun	Own pistol
Robbery	-0.007 (0.038)	-0.016 (0.034)
Arrest	0.007 (0.018)	0.013 (0.018)
Fear	0.006 (0.013)	-0.041 (0.012)
Predicted victim	-2.5 (0.064)	-0.006 (0.027)
Predicted arrest	-0.061 (0.034)	-0.348 (0.051)
Mean ownership	0.413 (0.036)	0.912 (0.144)
Mean ownership ^a	0.19 (0.092)	-0.46 (0.316)
Mean ownership ^b	0.29 (0.053)	-0.80 (4.0)
Mean ownership ^{a,b}	0.195 (0.147)	1.381 (0.293)
Mean ownership of high-school dropouts	0.010 (0.050)	0.402 (0.142)
Approve of hitting in retaliation	0.044 (0.011)	0.033 (0.01)
Confidence in courts	-0.023 (0.009)	-0.029 (0.009)
Police per square mile	-0.012 (0.008)	-0.011 (0.006)
Believe public officials care	-0.015 (0.008)	-0.022 (0.007)
Waiting period	0.005 (0.06)	0.019 (0.058)
Arrest	0.026 (0.021)	0.030 (0.035)
Waiting period × arrest	-0.059 (0.037)	-0.055 (0.035)

Notes: Regressions include all variables in Table 1. Sample sizes range from 6,000 to 19,000. The pistol regressions exclude respondents who own rifles or shotguns. The first three rows refer to a single regression, and the last three rows refer to a single regression. All other rows refer to distinct regressions. "Confidence in courts" ranges from 0 to 1 with higher values indicating more confidence in the Supreme Court. "Police per square mile" within the state regressions includes a control for state population density. "Believe public officials care" ranges from 0 to 1, where higher values indicate a belief that most officials care about the average man.

^a Regressions include state fixed effects.

^b Gun ownership or cohort is instrumented for with demographics.

fixed effects, the results remain for overall gun ownership but the pistol ownership result disappears. The following row addresses the possibility that omitted variables drive this

correlation by instrumenting for peer-group gun ownership using the level of gun ownership predicted by the basic variables of these peers. The last "mean ownership" row uses both instruments and includes state fixed effects. Overall, we take our results to mean that there is a sizable strategic complementarity in gun ownership: individuals who are surrounded by gun-owners also want guns themselves. These results also support the hypothesis that these effects will be most important among visible guns, not pistols.

To test whether this effect reflects self-protection against criminals, we look at the effect on ownership among high-skill persons of mean gun ownership among high-school dropouts. Gun ownership is explained by peer-group gun ownership, not by high-school-dropout gun ownership. However, pistol ownership is influenced by pistol ownership among the less skilled, supporting the view that people who just own pistols are motivated more by self-protection against crime.

The next row in the table shows that individuals who answer yes to the question "Would you approve of hitting someone who hit your child?" are more likely to own guns. Gun ownership appears to be associated with a general taste for violent retribution. The next row shows that gun ownership is negatively correlated with confidence in the Supreme Court. Gun ownership is also negatively correlated with the number of police per square mile in the state, holding overall population density constant. In the fourth row from the bottom, we see that gun-owners are less likely to believe that public officials care about them, suggesting that private and public justice appear to be substitutes.

The final three rows in the table examine the effects of waiting periods. These estimates include state fixed effects and are identified exclusively from changes in state laws. Waiting periods do not seem to reduce gun ownership generally, but they do reduce gun ownership among people who have been arrested, relative to the overall population. We found no connection between any other gun-control laws and the actual amount of gun ownership.

Gun-owners appear to be neither criminals nor victims, but rather individuals who are members of social groups where gun owner-

ship is the norm. Gun ownership appears to become a social norm when there is mistrust of public justice or where there is a tradition of private retribution.

REFERENCES

- Canada, Geoffrey. *Fist, stick, knife, gun: A personal history of violence in America*. Boston, MA: Beacon, 1995.
- Cook, Philip and Moore, Mark. "Gun Control." *Annals of the American Academy of Political and Social Science*, 1981, 455(2), pp. 267-94.
- DiPasquale, Denise and Glaeser, Edward L. "Incentives and Social Capital: Are Homeowners Better Citizens." *Journal of Urban Economics*, 1998 (forthcoming).
- Lott, John R. *More guns, less crime*. Chicago: University of Chicago Press, 1998 (forthcoming).
- Nisbett, Jon and Cohen, Dov. *Psychology of violence in the South*. Boulder, CO: Westview, 1996.

Guns, Violence, and the Efficiency of Illegal Markets

By JOHN J. DONOHUE III AND STEVEN D. LEVITT*

In economics, the standard mechanism for allocating scarce resources is the market. A smoothly functioning market, however, is built upon legally enforceable contracts and property rights. In the absence of law, it is likely that violence (or the threat thereof), rather than prices, is the means by which resources will be allocated. Interactions among animals provide clear evidence for this claim. Dominance hierarchies based on fighting ability, also sometimes known as pecking orders, have been documented across a wide variety of species (e.g., primates, chickens and other birds, reptiles, lobsters) and a broad range of resources including food, nesting sites, and access to mates (Warder C. Allee, 1938; John Alcock, 1993). Evidence suggests that violence also plays a critical role in human interactions when property rights are not legally enforceable (e.g., drug dealing and extortion) (see e.g., Peter Reuter, 1983; Geoffrey Canada, 1995).

In this paper, we analyze the determinants of the efficiency with which illegal markets allocate scarce resources. We develop a stylized model in which players compete for a fixed prize, with the winner determined by fighting ability. Efficiency in this context is determined by the amount of resources spent on fighting. Two factors affecting efficiency emerge from the model: *lethality* and *predictability*. Perhaps surprisingly, the use of more lethal mechanisms for resolving disputes does not have a clear impact on the social costs of violence. The intuition underlying this result is that, as the costs of losing a fight rise, the

willingness to fight falls. We show that holding other factors constant, the resources spent on fighting are lowest when the cost of losing is either very low or very high (e.g., nuclear deterrence), but over a wide range of lethality levels, the overall social costs of fighting are fairly stable.

In contrast, the costs of violence are critically linked to the predictability of dispute outcomes (i.e. the certainty with which potential combatants know who will be victorious *ex ante*). When the outcome of a conflict is highly correlated with observable characteristics such as strength or size, there is little need to actually fight. Thus unpredictability, all else equal, increases the expected payoff to fighting for the lower-ranked member, leading to more conflicts.

I. The Formal Model

In this section we formalize the intuition of the preceding discussion using a simple model that omits a number of potentially important considerations (e.g., private information and dynamic reputation effects). Nonetheless, the model provides a reasonable starting point for thinking formally about the issues at hand.

The structure of the model, which shares many common characteristics with the tournaments literature (Edward Lazear and Sherwin Rosen, 1981), is as follows. There are two players.¹ Each player i takes exactly one action, a_i , a decision about whether or not to fight; that is, $a_i \in \{\text{fight, no fight}\}$. The players are competing for a single prize which provides a payoff W to the winner. In order to

* Donohue: Stanford Law School, Crown Quadrangle, Stanford, CA 94305-8610; Levitt: Department of Economics, University of Chicago, 1126 East 59th Street, Chicago, IL 60637, and the American Bar Foundation. We thank Susan Albers, Edward Glaeser, Tim Groseclose, Bruce Kobayashi, John Lott, Sherwin Rosen, Chris Snyder, and especially James Heckman for insightful comments and suggestions. Financial support of the National Science Foundation is gratefully acknowledged.

¹ Given the functional forms we have adopted, the model readily expands to accommodate any finite number of players, although closed-form solutions become difficult to obtain. Space constraints preclude a detailed derivation of the N -player game, but we present simulation results from such models later in the paper. The model is equally applicable to individuals or groups of individuals, such as competing gangs.

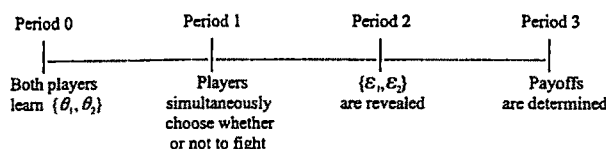


FIGURE 1. THE TIMING OF THE GAME

be eligible to win the prize W , a player must fight. If a player fights and loses, he receives a payoff $-C < 0$. Players who elect not to fight receive a default payoff normalized to zero.

Each player is characterized by a fighting ability F_i , where

$$(1) \quad F_i = \theta_i + \varepsilon_i.$$

What we will hereafter refer to as the observable component of fighting ability, $\{\theta_1, \theta_2\}$, is common knowledge to both players; $\{\varepsilon_1, \varepsilon_2\}$, on the other hand, is unobservable, even to the player himself (i.e., player i does not know ε_i). This unobservable component can be thought of as randomness in fight outcomes. The θ 's are independently and identically distributed normal with mean zero and variance equal to $0.5\sigma_\theta^2$. The ε 's are assumed to be independently and identically distributed with a type-1 extreme-value distribution,² characterized by

$$(2) \quad \Pr[\varepsilon_i \leq \varepsilon] = \exp[-\exp(-\varepsilon/\sigma_\varepsilon)].$$

This distribution proves to be extremely tractable, as will become apparent. Visually, the type-1 extreme-value function resembles a normal distribution, but with a thick right tail. The σ_ε^2 term influences the dispersion of the distribution.

The timing of the game is as shown in Figure 1. In period 0, the observable components of fighting ability (the θ 's) become common knowledge. Based on this symmetric (but incomplete) information on fighting abilities, the players simultaneously choose whether or not to fight. After each player decides whether or not

to enter the fight, the unobservable components of fighting ability $\{\varepsilon_1, \varepsilon_2\}$ are revealed, and the winner is determined. The winner is the player with the highest value of F_i among the set of players who elected to fight in period 1. If only one player chooses to fight, he automatically wins the prize W . If neither player opts to fight, no prize is awarded.

It is immediately evident that equilibrium must involve at least one player choosing to fight. For those equilibria involving one player choosing to fight and the other electing not to fight, no fight occurs and there are no resources expended on fighting.³ Only when both players elect to fight will a fight take place.

Let P_i equal the probability that player i wins the prize W conditional on both players choosing to fight:

$$(3) \quad P_i = \Pr(W_i | \theta_1, \theta_2, \sigma_\varepsilon) \\ = \Pr(\theta_i + \varepsilon_i > \theta_j + \varepsilon_j).$$

In general, there is no simple numerical solution to the relationship in equation (3). It has been shown, however, that if two independent random variables each have the same type-1 extreme-value distribution, then their difference has a logistic distribution (Norman Johnson and Samuel Kotz, 1970; Domencich and McFadden, 1975). It is this result that motivates our earlier distributional assumptions. Consequently, P_i can be expressed as

$$(4) \quad P_i = \frac{\exp\left(\frac{\theta_i}{\sigma_\varepsilon}\right)}{\exp\left(\frac{\theta_1}{\sigma_\varepsilon}\right) + \exp\left(\frac{\theta_2}{\sigma_\varepsilon}\right)}.$$

Given P_i , player i chooses to fight if and only if the expected payoff to fighting is

² Thomas Domencich and Daniel McFadden (1975) refer to this distribution as "Weibull (extreme value, Gnedenko)." We thank James Heckman for suggesting this functional form.

³ Which of the two players chooses to fight in such equilibria will depend not only on the parameters of the model, but also on player beliefs. Equilibria involving both players fighting (our primary focus) will not depend on player beliefs.

greater than the default payoff of not fighting, or mathematically,

$$(5) \quad WP_i - C(1 - P_i) \geq 0.$$

In order for a fight to occur, both players must satisfy equation (5). Noting that $P_1 + P_2 = 1$, the conditions for a fight can be written as

$$(6) \quad \frac{P_1}{P_2} \geq \frac{C}{W} \quad \text{and} \quad \frac{P_2}{P_1} \geq \frac{C}{W}.$$

Substituting equation (4) into equation (6), taking logs, and rearranging yields

$$(7) \quad \theta_1 - \theta_2 \geq \ln\left(\frac{C}{W}\right)\sigma_\varepsilon$$

$$\theta_1 - \theta_2 \leq -\ln\left(\frac{C}{W}\right)\sigma_\varepsilon.$$

The first expression is the cutoff for player 1's willingness to fight; the second expression is the threshold for player 2. The larger the cost associated with fighting and losing relative to the prize, the less willing players are to fight. If $C > W$, then those two conditions can never be simultaneously satisfied, and there will never be a fight. Intuitively, $C > W$ means that the combined expected payoff given that a fight occurs is negative, implying that at least one player must have an expected payoff that is negative. As long as $C < W$, the log term is negative, allowing for the two conditions to be simultaneously satisfied. The greater is the uncertainty in the determination of the fight outcome (i.e., the greater is σ_ε), the higher the chances that the player with the lower observable fighting ability will be victorious. Since it is always the weaker player who represents the binding constraint on a fight occurring, less predictability of fight outcomes will lead to more fights.

The expression in equation (7) is conditional on the particular values of θ that are observed. In order to make general statements about the probability of fights, one must integrate over the joint distribution of θ_1 and θ_2 . Because of the normality assumption for θ , the difference between the two θ 's is itself normally distributed. Figure 2 captures pictorially the likelihood that a fight occurs. The

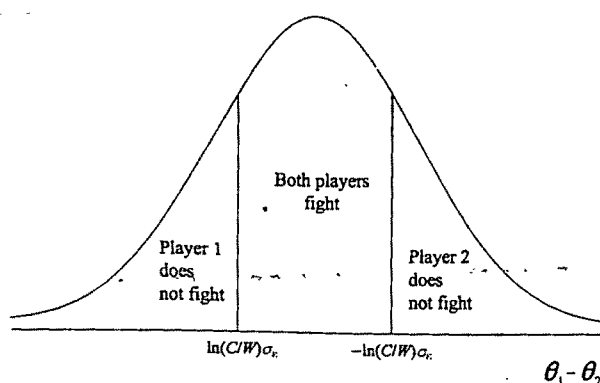


FIGURE 2. THE CONDITIONS UNDER WHICH A FIGHT OCCURS

probability density function of $\theta_1 - \theta_2$ is pictured in the figure. The symmetric vertical lines to the left and right of center represent the cutoff points below which player 1 and player 2 respectively choose not to fight. The middle area between the two lines represents the probability that a fight will occur. Increasing C , lowering W , or reducing σ_ε will shift the vertical lines inward, reducing the number of fights. Increasing the variance of $\theta_1 - \theta_2$ reduces the mass in the middle area, also reducing the number of fights.

Noting that the variance of $\theta_1 - \theta_2$ is σ_θ^2 , the unconditional probability that a fight occurs can be expressed as

$$(8) \quad \Pr(\text{Fight occurs}) = 1 - 2\Phi\left[\frac{\sigma_\varepsilon}{\sigma_\theta} \ln\left(\frac{C}{W}\right)\right]$$

where Φ represents the cumulative distribution function of the standard normal distribution. The term after the minus sign is the weight outside the vertical lines in Figure 2. Equation (8) provides an extremely convenient and intuitive characterization of the likelihood of a fight: (i) the relative importance of unobserved factors in determining fight outcomes (the ratio of the σ 's), which we will call *predictability*, and (ii) the log ratio of the costs of losing relative to the prize for winning, which we term the *lethality* of the fighting technology.

It is not simply the number of fights that matters, however, but rather the amount of resources expended in fighting that is of primary importance. Thus, we consider a variable

V (for violence) which is the expected value of the costs associated with fighting, obtained by multiplying equation (8) by C :

$$(9) \quad V = C \left(1 - 2\Phi \left[\frac{\sigma_\varepsilon}{\sigma_\theta} \ln \left(\frac{C}{W} \right) \right] \right).$$

Understanding the relationship between the costs of violence V and the parameters it depends upon is best accomplished visually. Figure 3 presents a graph of V as a function of the lethality and predictability of fight outcomes. Throughout the graph, W is held fixed at 100. Moving from left to right, C is allowed to vary from 0 to 100. The three lines traced out on the graph represent three different values of predictability. The curve labeled "most predictable" has a ratio of $\sigma_\varepsilon/\sigma_\theta = 0.1$. The "somewhat predictable" curve has a ratio of 0.5, and the "least predictable" curve has a ratio equal to 1. Moving along a given curve (i.e. holding predictability constant), rising lethality initially increases the costs of violence but then lowers it. At the two extremes, there are no costs of violence. When $C = 0$, fighting carries no costs; when $C = W$, fights never occur. The downward-sloping part of the curve underlies the logic of nuclear deterrence.

Comparison across curves demonstrates that predictability is strongly related to the costs of violence. For middle ranges of lethality, the costs of violence are roughly 25 percent of the prize being fought over in the least predictable case, but only 3 percent in the most predictable case. When players have better information *ex ante* about who will emerge victorious, the number of fights is lower.⁴

The same patterns observed in Figure 3 emerge more strongly as the number of players in the game increases. Space constraints preclude a full accounting of the N -player game. It is worth reporting simulation results for a five-player version of the game with $W = 100$, which we have described more fully in Donohue and Levitt (1997). While the curves traced in that game rose and fell with lethality as in Figure 3, the most striking feature of the simulations was that over a wide range of values for lethality ($20 < C < 80$), there

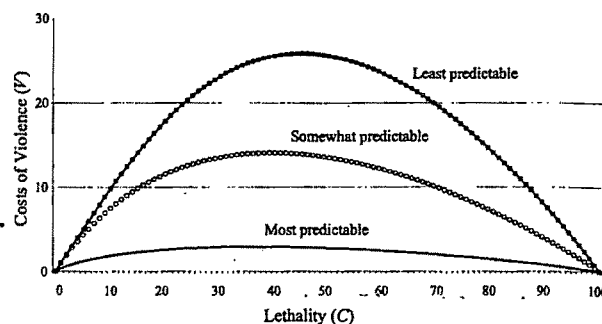


FIGURE 3. THE IMPACT OF LETHALITY AND PREDICTABILITY ON THE COSTS OF VIOLENCE

was virtually no relationship between lethality and violence. The key difference between the N -player game and the two-player game is the possibility of more than one fight in the former. This both mutes the sensitivity of the costs of violence to lethality and raises the overall costs of violence. Also, in the limit as the outcome of fights becomes completely unpredictable, the costs of fighting will rise to the point where all of the surplus associated with winning the prize is competed away. This result is similar to the destruction of surplus that occurs in a variety of preemption games, such as the timing of adoption of a new innovation (Drew Fudenberg and Jean Tirole, 1991).

II. Guns, Drug Markets, and Violence

We can now use the above model to examine reasons for the doubling of the juvenile homicide rate in the years from 1985 to 1995, in a period when the adult homicide rate declined slightly. The increase in juvenile homicides appears to coincide with two factors (Alfred Blumstein 1995): (i) a dramatic increase in drug distribution by street gangs, particularly crack cocaine, and (ii) a great rise in gun-carrying among juveniles, particularly for those involved in the drug trade. Virtually all of the increase in juvenile homicides over this time period is attributable to a rise in gun-related deaths.

In the notation of our model, the profits of the drug trade increased W (which tends to increase fighting), and the accompanying increase in gun usage increased C or lethality. The standard argument concerning the link between juvenile homicide rates and guns has focused on lethality. It is frequently said that

⁴ If players are risk-averse, then this result will be attenuated.

juveniles have always fought, but now they die because they fight with guns. As the model above makes clear, that argument is potentially flawed because it ignores the fact that more lethal weapons should lead participants to show greater discretion in their willingness to fight. Had there been a very lethal fighting technology adopted, but one for which the outcome was nonetheless highly predictable (e.g., the winner of a fight is first determined through fisticuffs, and then the loser is immediately executed), the number of violent deaths is unlikely to have increased so dramatically.

Our model suggests that the standard explanation for the link between guns and juvenile violence is inadequate because it ignores a critical factor: the unpredictability of dispute outcomes when juveniles arm themselves with guns. When fights involve less lethal weapons such as knives, observable factors (e.g., the physical appearance of the opponent, past fighting record, or number of people in the opposing gang) provide a good indicator of who will win the fight. With the introduction of guns, however, the factors that predict victory (e.g., lack of respect for human life, disutility of going to prison, high discount rate) are less observable, more variable over time, and subject to strategic manipulation. Guns are an equalizing force that makes the outcome of any particular conflict difficult to predict. All else held constant, this increases the willingness to fight among weaker combatants, leading to greater levels of violence.

III. Conclusions

This paper has examined violence as a mechanism for allocating scarce resources in a nonmarket setting. We demonstrate that the efficiency with which resources are allocated in that context are strongly positively related to the predictability of fight outcomes. The lethality of the weapons used, in contrast, has an indeterminate impact on the costs of violence, except at very low or very high levels of lethality. Our results suggest that the observed link between guns and homicide rates may not be primarily attributable to the lethality of guns, but rather to the lack of *ex ante* predictability of the winner when guns are involved in a fight. To paraphrase an often heard statement, "guns don't kill people, the unpredictability of guns kills people."

Three limitations to the arguments presented in this paper are important to emphasize. First, the model presented is directly relevant only to disputes over scarce resources carried out by rational actors. Many violent deaths are the result of arguments between spouses or as a consequence of suicide. In such circumstances, the strategic aspects that are central to the model of this paper may be less applicable. Second, we do not consider costs of violence that are external to the disputants. The introduction of guns may lead to the death of innocent bystanders caught in the cross-fire. Finally, our model does not attempt to explain what mechanism for dispute resolution is adopted. Given that the mechanisms used appear to vary substantially across time and space, endogenizing that choice would appear to be a useful extension of the current model.

REFERENCES

- Alcock, John. *Animal behavior: An evolutionary approach*. Sunderland, MA: Sinauer, 1993.
- Allee, Warder C. *Social life of animals*. New York: Norton, 1938.
- Blumstein, Alfred. "Youth Violence, Guns, and the Illicit-Drug Industry." *Journal of Criminal Law and Criminology*, Fall 1995, 86(1), pp. 10-36.
- Canada, Geoffrey. *Fist, stick, knife, gun: A personal history of violence in America*. Boston, MA: Beacon, 1995.
- Domencich, Thomas and McFadden, Daniel. *Urban travel demand: A behavioral analysis*. New York: Elsevier, 1975.
- Donohue, John and Levitt, Steven. "Guns, Violence, and the Pecking Order." Unpublished manuscript, University of Chicago, 1997.
- Fudenberg, Drew and Tirole, Jean. *Game theory*. Cambridge, MA: MIT Press, 1991.
- Johnson, Norman and Kotz, Samuel. *Continuous univariate distributions*. Boston, MA: Houghton-Mifflin, 1970.
- Lazear, Edward and Rosen, Sherwin. "Rank Order Tournaments as Optimum Labor Contracts." *Journal of Political Economy*, October 1981, 89(5), pp. 841-64.
- Reuter, Peter. *Disorganized crime: Illegal markets and the Mafia*. Cambridge, MA: MIT Press, 1983.

Lives Saved or Lives Lost?

The Effects of Concealed-Handgun Laws on Crime

By HASHEM DEZHBAKHSH AND PAUL H. RUBIN*

The role of handguns in crime has been the subject of extensive policy and academic debate in recent years. The interest in the issue has grown with the enactment of the restrictive (federal) Brady Bill and the adoption by many states of the right-to-carry concealed-handgun laws. These "shall issue" laws make it much easier for noncriminals to obtain licenses to carry concealed handguns. (Ten states passed such laws from 1977 to 1992, and 13 states have passed such laws since 1992.) This observed dichotomy in policy reflects the lack of consensus among policymakers regarding the role of handguns in violence. A similar disagreement exists in academic circles. For example, results reported in Philip Cook et al. (1995), Arthur Kellermann et al. (1995), Cook and Jen Ludwig (1996), and David Hemenway (1996) imply that an increase in gun ownership increases rates of crime. Daniel Polsby (1995) and John Lott and David Mustard (1997) are doubtful of such implications.

There are two theoretical possibilities regarding the effects of these laws. Increased gun ownership might lead to increased crime because of the increased availability of guns—the facilitating effect. Alternatively, because increased gun ownership might help potential victims to arm and protect themselves, thus increasing the criminals' uncertainty regarding an armed response, such laws might lead to reduced crime against persons—the deterrent effect. Which effect dominates might depend on population characteristics in a given jurisdiction. Concealed-handgun laws might

lead to increases in crime in some jurisdictions, while leading to decreases in other jurisdictions.

In a controversial paper, Lott and Mustard (1997) have examined this issue. They find that passage of concealed-handgun (shall-issue) laws by a state causes a significant reduction in violent as well as property crimes (Lott and Mustard, 1997 table 11). They attribute the results to a deterrent effect: as criminals become aware that victims might be armed, they reduce the rate of commission of crimes.¹

We believe Lott and Mustard's findings are suspect, mainly because of the way they parameterize and measure the effect of permissive handgun laws on crime. They model the effect as a shift in the intercept of the linear crime equation they estimate at the county level. This approach is predicated on two assumptions: (i) all behavioral (response) parameters of this equation (slope coefficients) are fixed (unaffected by the law), and (ii) the effect of the law on crime is identical across counties. Obviously, if the law affects the behavior of the criminals or of citizens, then these response parameters should change, and not only the intercept term. Moreover, it seems highly unlikely that the magnitude of the ef-

¹ The Lott-Mustard results that have received the most attention from media and from critics, and that the authors themselves emphasize in the abstract, are based on their table 3, which uses least squares and does not correct for the endogeneity of the arrest rate. This is the specification that shows a trade-off between violent and property crimes as states adopt "shall issue" laws. In the two-stage least-squares regression reported in Lott and Mustard (1997 table 11) the trade-off does not exist; they find a significant, and very substantial, decrease in all categories of crime in this specification. Our results are also based on two-stage least-squares methods. We should also note that the Lott-Mustard results in their table 11 are based on predicted arrest rates, from the first-stage estimation, which does not include the county fixed-effect estimates. This may also render suspect the results in their table 11.

* Department of Economics, Emory University, Atlanta, GA 30322-2240. We thank John Lott and David Mustard for providing us with the data and Bill Sribney, from Stata Corporation, for helpful programming suggestions. We also acknowledge financial support from the Emory University Research Committee. The usual disclaimer applies.

fects such laws may have on crime rates in a county would be independent of economic and demographic characteristics of the county. In fact, the effect may vary with the age and gender composition of the population and the economic conditions of the counties, among other things. Others who have commented on Lott and Mustard (1997) (e.g., Ludwig, 1996; Dan Black and Daniel Nagin, 1998) have overlooked this problem, and consequently their alternative estimates may be subject to the same criticism; see also the reply by Lott (1998) to Black and Nagin's criticisms.

We reexamine the effect of permissive concealed-handgun laws on crime using a procedure to overcome these shortcomings, allowing all behavioral parameters of the model to change. Our method also allows the effect of the law on crime rates to be heterogeneous across counties so that we can infer how various factors influence the magnitude of the change in crime resulting from such laws. More specifically, we project the 1992 crime rate for counties without such laws if they had adopted the law by 1992. We then compare these projections, which are a function of county characteristics, with actual crime data for 1992 to infer how the absence of the law has affected crime in these counties. We also examine the relationship between these projected changes and county characteristics.

I. Model and Estimation Approach

Lott and Mustard (1997) use county-level panel data to estimate several linear crime equations. The dependent variable is one of several crime rates: murder, rape, aggravated assault, robbery, burglary, larceny, and auto theft. The regressors include the corresponding arrest rate, a host of economic and socio-demographic factors, and a (0 or 1) binary variable measuring the status of the concealed-handgun law. The other regressors serve as control variables. The model they estimate is therefore

$$(1) \quad C_{it} = \alpha + \gamma I_{it} + \beta A_{it} + \delta X_{it} + \varepsilon_{it}$$

where I is the binary variable, A is the arrest rate, X includes the economic and demo-

graphic variables and a set of time and county dummies (one for each sampling year or county), ε is the regression error, and i and t denote counties and time periods, respectively.

Lott and Mustard's inference about the effect of concealed-handgun laws on various categories of crime is based on the sign and statistical significance of the estimated coefficient of the binary variable: the estimate of γ . A positive and significant estimate suggests that concealed-handgun provisions would increase the crime rate, while a negative and significant estimate points to the contrary conclusion. This representation assumes that the law only affects the intercept of the crime equation, so the crime equations for the counties with and those without the law have different intercepts but identical slopes. The coefficient of the binary variable captures the difference between the two intercept terms.

The implicit assumptions behind this characterization is that all other response parameters are identical for the two groups of counties and that all counties that adopt the law will observe identical changes in their crime rates. Both assumptions are unwarranted. Each of these parameters measures the change in the crime rate resulting from a change in the corresponding control variable (arrest rate, economic variables, or demographic characteristics). If adopting the law does indeed affect the action of criminals or their potential victims, it would do so by altering these response parameters.² The statistical consequences of ignoring such changes are biased estimation and potentially invalid results. Moreover, assuming the effect of the law on crime rates (parameter γ) to be fixed across counties is unjustified given the diversity of the county characteristics. Finally, using an intercept change to measure the effect of a change in the law causes the estimated effect to be fragile with respect to small specification changes. William Bartley et al. (1998) use

² Lott and Mustard (1997) in one of their least-squares regressions (their table 7) which, as explained earlier, is inappropriate in the present context, allow the slope coefficient of county population (one of the 53 regressors in their model) to change. However, they do not examine the combined effect.

extreme-bounds analysis to examine the range of the estimates of the intercept change and find it to be apparently quite wide in many cases.

We reexamine the effect of the concealed-handgun laws on crime using a regression-based procedure that overcomes these limitations. We allow all behavioral parameters of the model to respond to a change in the status of the law. The data will then show which of these parameters the law indeed affects. We implement this parameter flexibility by estimating two separate crime equations, one for counties in states with a concealed-handgun law and the other for the remaining counties:

$$(2a) \quad C_{L,it} = \alpha_L + \beta_L A_{L,it} + \delta_L X_{L,it} + \varepsilon_{L,it}$$

$$(2b) \quad C_{NL,it} = \alpha_{NL} + \beta_{NL} A_{NL,it} + \delta_{NL} X_{NL,it} + \varepsilon_{NL,it}$$

where the subscripts L and NL indicate the presence or the absence of the concealed-handgun law, respectively.

First, we examine whether the law affects the response parameters by using an asymptotic Wald test of the null hypothesis $H_0: \Theta_L = \Theta_{NL}$ against the alternative $H_0: \Theta_L \neq \Theta_{NL}$, where Θ denotes (β, δ) .³ A rejection of the null hypothesis implies that the law affects the response parameters of the model and therefore the crime rate. Following Isaac Ehrlich (1973), in all our estimations we treat the arrest rate, A , as an endogenous variable that is affected by such variables as the lagged crime rate, economic and demographic variables in the crime equation, police employment and payroll, and variables to control for political influences. These include percentage of Republican presidential votes and the percentage of a states' population who are members of the National Rifle Asso-

ciation (NRA); see Lott and Mustard (1997 p. 42). We estimate equations (2a) and (2b) along with the corresponding arrest equations via two-stage least-squares (2SLS) analysis, allowing the concealed-handgun law to shift the coefficients of the arrest equation in the first stage of estimation; such shifts are incorporated in cases where the Wald test applied to an arrest equation suggests that such a change is warranted. This ensures the consistency of the second-stage estimates. In all our estimations, we correct the residuals from the second-stage least square to account for using predicted rather than the actual arrest rate in estimation of the crime equation (see e.g., Russell Davidson and James G. MacKinnon, 1993 Ch. 7).

We estimate for each county the direction and extent of the change in crime rate that may result from introducing the concealed-handgun law. We determine how different the crime rate would have been during 1992 in the counties that did not have the concealed-handgun law in place, had they adopted the law by 1992. We obtain these estimates, which are useful for policy purposes, simply by switching the estimates of the behavioral parameters in equations (2a) and (2b) and computing the resulting predicted values for the dependent variable (the crime rate) over the relevant year. The estimates are obtained from $\hat{C} = \hat{\alpha}_L + \hat{\Theta}_L Z_{NL}$, where Z_{NL} denotes the regressors in equation (2b). These are predicted crime rates conditional on adopting the concealed-handgun law. The difference between these predicted crime rates and the actual crime rates is our measure of the effect of concealed-handgun laws on crime. We emphasize that our interest is in estimating the expected 1992 crime rates conditional on the law being in place in a county that did not have it in 1992. This estimate is then compared with the county's actual 1992 crime rate to estimate the expected change. It is important to note that, in the above comparison, one should not use the county's predicted 1992 crime rate in the absence of the law, $\hat{\alpha}_L + \hat{\Theta}_L Z_{NL}$, in place of the observed crime rate. This is because the former has no useful information for our inference that is not contained in the county's observed 1992 crime rate. Therefore, if we used $\hat{\Theta}_{NL} Z_{NL}$ instead of

³ The Wald statistic is the quadratic form constructed on the estimate of the difference $(\Theta_L - \Theta_{NL})$. The statistic is asymptotically distributed as a χ^2 variate with degrees of freedom equal to the number of parameters tested (Leslie Godfrey, 1988 Ch. 4; Andrew Lo and Whitney Newey, 1985).

the actual crime rate, we would add extra noise (residual), thus reducing the accuracy of the inference. Also, note that all the information relevant to adopting the law is incorporated in $\hat{\Theta}_L$, which is estimated using counties with the law.

We summarize the predictions we obtain to make inferences about the scope of the potential influence of the law in each state that did not have a concealed-handgun law in place in 1992. The projections are then further analyzed to determine factors that influence their direction or magnitude. The effect of concealed-handgun laws, therefore, may vary with population density, racial and gender characteristics, income, and so forth.

II. Data and Results

We use the data provided by Lott and Mustard (1997). The complete data set covers 3,054 counties for the period 1982–1992. The data set includes the FBI's crime data for murder, rape, aggravated assault, and robbery which comprise "violent crime" and auto theft, burglary, and larceny which comprise "property crime." The series also include the corresponding arrest rate for these nine crime categories,⁴ population and population density, population characteristics for 36 age and race segments (black, white, and other; male and female; and age divisions), retirement payments per person over 65 years of age, and the ratio of real per capita income to personal income, unemployment insurance payments, and income maintenance payments. The data set also includes state-level data on police employment and payroll, the percentage of Republican presidential votes, and the percentage of each state's population who are members of the NRA. This is the most exhaustive panel data set available for research in this area.

Using a Wald test for all nine categories of crimes, we find that the Lott-Mustard hypoth-

esis of no change in slope coefficients is rejected strongly (with p values close to zero) for all categories of crime.⁵ That is, in all cases, there are significant changes in the slope coefficients, so that assuming all changes to be embedded in the intercept is incorrect. This suggests that the Lott-Mustard results are biased due to a misspecification. Similar results for the arrest equation, used in the first stage of the 2SLS estimation, indicate that the coefficients of these equations also change with the law. In fact, we incorporate these changes when obtaining the predicted arrest rates. A comparison of our predicted arrest rates to those of Lott and Mustard (1997) reveals the inaccuracy introduced by limiting the change to the intercept term. For example, depending on the crime category, the mean-square errors of Lott and Mustard's predicted arrest rates are 1.5–5.2 times larger than those of ours. Their predicted arrest rates also include a large number of negative values (i.e., out of about 33,000 observations for each crime category, over 19,000 are negative for auto theft, 9,900 for aggravated assault, and 13,500 for property crimes).⁶

We use the two-stage procedure described earlier to estimate the hypothetical effect on crime in each county in states that did not have a concealed-handgun law in place if such a law had been in effect in 1992. We examine these effects in two ways, both on a county-by-county basis. First, we examine for each crime and county the predicted effect of changing the law. Second, we examine the effect of county characteristics on predicted change in crime rates for each aggregated crime category.

Table 1 contains summary statistics derived from these county-level conditional predictions; more extensive results including estimated percentages are available from the

⁴ The arrest series we use are slightly different than those used by Lott and Mustard (1997). As indicated in the last paragraph of their data appendix, Lott and Mustard use arrest series that erroneously contain zero's instead of some missing values. We use the corrected series.

⁵ Results reported here are slightly different than those in an earlier draft, where we did not use population weights in estimation. Here the weights are included to make our specification as close to that of Lott and Mustard (1997) as possible, given the difference in the way we model the effect of the law.

⁶ These large negative values may be partly due to Lott and Mustard's omission of the county fixed effects from the predicted arrest rates.

TABLE 1—THE PREDICTED EFFECT OF ADOPTING
CONCEALED-HANDGUN LAWS ON VIOLENT CRIMES
IN STATES WITHOUT SUCH LAWS IN 1992

State	Murder	Rape	Robbery	Aggravated assault
Arizona	no	+	+	+
Arkansas	no	—	no	—
Colorado	no	no	no	+
Illinois	—	no	mix	mix
Iowa	no	mix	+	mix
Kansas	—	—	+	+
Kentucky	—	mix	+	—
Louisiana	no	no	+	—
Maryland	no	no	+	mix
Michigan	no	no	+	no
Minnesota	—	no	+	mix
Missouri	no	mix	mix	mix
Nebraska	no	no	no	mix
Nevada	no	no	+	no
New Jersey	no	no	no	+
New Mexico	no	no	+	+
New York	no	no	no	mix
North Carolina	no	+	no	+
Ohio	—	mix	+	mix
Oklahoma	no	no	no	mix
South Carolina	no	+	no	no
Tennessee	no	mix	+	mix
Texas	—	—	mix	mix
Utah	no	no	no	+
Wisconsin	no	+	+	—

Notes: Entries indicate the effect of such laws on crimes in 1992 for each state, had the state adopted the law by then. A + (—) indicates an unambiguous increase (decrease), "no" indicates that no county would have been affected, and "mix" indicates increases for some counties in the state and decreases for others. The states with "no" in all four categories are dropped for brevity; these include Alaska, California, Delaware, the District of Columbia, Hawaii, Massachusetts, Rhode Island, Wyoming, and Pennsylvania (Philadelphia County).

authors upon request. Our results suggest that for counties in six states a concealed-handgun law would have reduced murder rates, and for all counties in the other 27 states it would have been ineffective. Overall, the results indicate a relatively small, and crime-reducing, effect of concealed-handgun laws on murder rates. There would have been little effect on rape, with 22 states unaffected, four states with unambiguous increases, and two states with unambiguous decreases. The effect on robbery would have been an increase in crime for many states. For counties in 13 states, there would have

been unambiguous increases in robbery; there would have been a mixed effect (increases in some counties and decreases in some) in only three states. The overall increase in robbery is not surprising given that concealed handguns add little deterrence in this case. Many potential robbery targets already have armed protection; therefore, concealed firearms would not increase the deterrence factor. It would, however, have a crime-facilitating effect, helping the robbers.

For aggravated assault, 12 states would have been unaffected, seven states would have been adversely affected, and four states would have observed a drop in crime. The result for the remaining states is mixed. For the three categories of property crime, the effect would have been more mixed. The largest percentage of counties predicted to be affected in one direction by changing the law would have been the 15 percent of counties predicted to experience an increase in larceny; all other predicted percentage changes in any direction are less than 10 percent.

We next determine which characteristics of counties are associated with increases or decreases in each aggregate type of crime (violent and property crime) for counties. We do this by regressing the predicted change in crime rates as a function of a list of demographic and economic variables for the county. The economic variables, all measured per capita, are personal income, unemployment insurance, and retirement payments per person over 65. We also include (predicted) arrest rates and population density. We include demographic variables. Since most crime is committed by young males, we include the number of males 10–29 years old, and similarly for females. We include persons 65 and over, who are perhaps more likely to be victims than perpetrators of crimes.⁷ Finally, we include per capita measures of number of NRA members in the state and police pay-

⁷ Experiments with other specifications indicate that this specification provides most of the useful information in the data and is sufficiently aggregated so that the results are easily interpreted.

roll. In all cases, we measure the effect of the relevant variable on predicted changes in crime by category of the passage of a concealed-handgun law by county.

Regression results show that for counties that spend relatively more on police the laws lead to crime reductions. This is plausible: higher spending on police will not affect the deterrent benefit of handguns but will reduce the facilitating effect of handguns that benefits criminal activity. On the other hand, for counties with higher arrest rates, passage of shall-issue laws leads to increased crime, perhaps because there are more criminals in these counties. The other consistent results are for likely victims: more elderly people and more young (10–29-year-old) nonblack females are associated with reduced crime as a result of passage of gun laws. This may represent evidence of the deterrent effect in some cases. Evidence that potential criminals (generally, young males) use these laws to obtain guns and commit crimes is weak; only one of the four possible coefficients is significant and positive, and one is negative (t values for both black and nonblack young males are less than 0.2 in the property-crime estimate).

III. Concluding Remarks

We have reexamined the effect of concealed-handgun laws on crime rates using a statistical procedure that overcomes the limitations of previous studies. Our procedure allows us to assess the full implications of the right-to-carry gun provisions. We find that the results of concealed-weapons laws are much smaller than suggested by Lott and Mustard (1997) and by no means negative across all crime categories. For murder, for example, there is only at best a small reducing effect. For robbery, many states experience increases in crime. For other crimes, results are ambiguous, with some counties showing predicted increases, and some predicted decreases.

We also examine demographic and other influences on the likely effect of passage of laws on crime rates. We find that there are predictable patterns on the effect of shall-issue laws on crime. For example, counties spending more on police could expect a decrease in

crime from the passage of a law or a smaller increase where the law leads to an increase in crime. The sort of analysis developed here could be used to enable policymakers to tailor laws more carefully to particular conditions in a jurisdiction. More research in this important area is warranted.

REFERENCES

- Bartley, William; Cohen, Mark and Froeb, Luke. "The Effect of Concealed Weapon Laws: An Extreme Bounds Analysis." *Economic Inquiry*, 1998 (forthcoming).
- Black, Dan and Nagin, Daniel. "Do 'Right-to-Carry' Laws Deter Violent Crime?" *Journal of Legal Studies*, January 1998, 27(1), pp. 209–19.
- Cook, Philip J. and Ludwig, Jen. "Guns in America: Results of a Comprehensive National Survey on Firearms Ownership and Uses." Report prepared for the National Institute of Justice, Washington, DC, 1996.
- Cook, Philip; Molliconi, Stephanie and Cole, Thomas. "Regulating Gun Markets." *Journal of Criminal Law and Criminology*, Fall 1995, 86(1), pp. 59–92.
- Davidson, Russell and MacKinnon, James G. *Estimation and inference in econometrics*. New York: Oxford University Press, 1993.
- Ehrlich, Isaac. "Participation in Illegitimate Activities: A Theoretical and Empirical Investigation." *Journal of Political Economy*, May–June 1973, 81(3), pp. 521–65.
- Godfrey, Leslie. *Misspecification tests in econometrics: The Lagrange multiplier principle and other approaches*. Cambridge: Cambridge University Press, 1988.
- Hemenway, David. "Survey Research and Self-Defense Gun Use: An Explanation of Extreme Overestimates." Working paper, Harvard University, 1996.
- Kellermann, Arthur; Westohal, Lori; Fischer, Lauri and Harvard, Beverly. "Weapon Involvement in Home Invasion Crime." *Journal of the American Medical Association*, 14 June 1995, 273(22), pp. 1759–62.

- Lo, Andrew and Newey, Whitney. "A Large-Sample Chow Test for the Linear Simultaneous Equation." *Economics Letters*, 1985, 18(4), pp. 351-53.
- Lott, John. "The Concealed Handgun Debate." *Journal of Legal Studies*, January 1998, 27(1), pp. 221-43.
- Lott, John and Mustard, David. "Crime, Deterrence, and Right-to-Carry Concealed Handguns." *Journal of Legal Studies*, January 1997, 26(1), pp. 1-68.
- Ludwig, Jens. "Do Permissive Concealed-Carry Laws Reduce Violent Crime?" Mimeo, Georgetown University, 1996.
- Polsby, Daniel. "Firearm Costs, Firearm Benefits, and the Limits of Knowledge." *Journal of Criminal Law and Criminology*, Fall 1995, 86(1), pp. 207-20.

Criminal Deterrence, Geographic Spillovers, and the Right to Carry Concealed Handguns

By STEPHEN G. BRONARS AND JOHN R. LOTT, JR.*

Increased law enforcement or penalties may deter crime, but they may also cause criminals to move to other crimes or other areas. When a car owner uses "The Club," auto thieves respond by moving on to other cars unprotected by such a lock.¹ Auto theft should decline if more car-owners lock their cars as the cost of searching for easily stolen cars rises, and the extent of the decline is an empirical question. The movement of criminals between jurisdictions is similar, though more difficult to predict. Greater law enforcement may cause criminals to migrate to bordering areas. However, if criminals were using the original location as a "staging" area for crime into surrounding communities, crime should decline elsewhere (e.g., clamping down on automotive "chop shops" or fences thus reduces the return to theft in surrounding communities).² Despite the potential importance for evaluating law enforcement, spillovers have not been investigated empirically. Existing studies of local law enforcement might either over- or underestimate the overall benefits from deterrence.

This paper examines whether the adoption of a shall-issue concealed-weapons law in one state alters crime in neighboring areas. The shall-issue law guarantees a citizen who meets

certain objective criteria the right to carry a gun. Lott and David B. Mustard (1997) found a strong local deterrent effect of the law, but we wish to examine whether much of this merely represents crime moving elsewhere (see also William Alan Bartley and Mark A. Cohen [1998] and Lott [1998a, b]). Alternatively, simultaneously passing shall-issue laws in a number of states might reduce a state's crime rate by even more than simply adopting the law on its own. By incorporating spillover effects into our analysis, we estimate the aggregate effect of shall-issues laws on crime rates. To our knowledge, this paper constitutes the first attempt to include spillover effects in estimating criminal deterrence. The methodology used here could be extended to other crime-deterrence policies.

I. Data and Methodology

This paper uses annual cross-section time-series county-level crime data for the continental United States from 1977 to 1992. We weigh all regressions by county population, and we include county-fixed effects and year-fixed effects in each of the regressions reported below. Therefore, all the estimated effects of concealed-weapons law on crime are derived from changes within a given county, relative to the year-to-year changes in the overall U.S. crime rate.

For each county-year we employ a set of demographic characteristics: the distribution of the county's population by age, race, sex, average income, welfare, and population density. We include this vector of demographic characteristics in all the regression characteristics below (see Lott and Mustard [1997] for further discussion of these variables). The crime data are from the FBI's Uniform Crime Report. Our primary dependent variables are the reported crime rates per 100,000 population per county per year for nine different

* Department of Economics, University of Texas, Austin, TX 78712, and School of Law, University of Chicago, Chicago, IL 60637, respectively. Lott received support from the John M. Olin Program in Law and Economics at the University of Chicago Law School.

¹ For a survey of the theoretical literature on protective spillovers see Kermit Daniel and Lott (1995).

² "Thriving" criminal communities may also facilitate people becoming criminals. If there are fences on every street corner in a city, it is much easier for thieves from outlying areas to come to the city to sell their stolen goods. Penalties that affect the number of "fences" (either by directly punishing the fences or reducing the number of criminals who seek them out) could reduce crime in surrounding areas.

crime categories: overall violent crimes, murders, rapes, robberies, aggravated assaults, overall property crimes, burglaries, auto thefts, and larcenies.

We extend the analysis of Lott and Mustard (1997) to incorporate the impact on neighboring counties. A neighboring county is here defined as any county with a geographic center within 50 miles of the geographic center of the home county.³ We compare variation in crime rates across counties within the same state over the same time period. A change in concealed-weapons laws in a given state is assumed to affect crimes in neighboring counties. For example, when Georgia introduced a shall-issue law in 1989, it likely affected neighboring counties in Florida, Tennessee, Alabama, and South Carolina. The effects of the Georgia law on neighboring counties in Tennessee was estimated by comparing time-series fluctuations in crime for these counties to other counties in Tennessee. Further, some states bordering Georgia had already passed these laws, and we assume that, for this reason, the spillover effects could also differ across states.

While our initial specifications correspond to Lott and Mustard's simplest estimates, our other specifications are different. First, we control for either the violent or property-crime arrest rate depending upon whether the crime rate being studied is that of violent or property crime. This mitigates any spurious relationship between crime and arrest rates that might arise because the arrest rate is a function of the crime rate. It also helps solve the missing-value problem arising because the arrest rate is undefined when no crime occurs,⁴ but it still allows us to control for changes over time in the effectiveness of law enforcement. Including the arrest rate for the detailed crime category or omitting the arrest rate from the

regression produces similar results. Second, we use the crime rate, rather than the log of the crime rate, as the dependent variable to test whether the results are sensitive to examining absolute rather than percentage changes in the crime rate.

Again, following Lott and Mustard, our regressions capture the effects of concealed-weapons laws over time by including linear trend variables for the years before and after the change in the law, along with a shall-issue dummy variable.⁵ Theoretically, the long-run impact ought to be larger as law-abiding citizens and criminals alike adapt to changes in the weapons laws. This simple parametric specification can accommodate a once-and-for-all change as well as cumulative changes which may reinforce or mitigate the immediate impact. For many of the jurisdictions in our sample only a few years elapse after passage of the law, which means that we cannot confidently evaluate the cumulative impact beyond four years.

II. Empirical Results

Part A of Table 1 reports murder, rape, and robbery regression results similar to those in Lott and Mustard's (1997) table 3, with the addition of spillover effects for shall-issue laws in neighboring counties (results that are not shown are available from the authors upon request). We use the log of the crime rate as the dependent variable. Independent variables include the arrest rate for the corresponding crime category, a dummy variable for a state's own shall-issue law, a variable measuring the (population-weighted) fraction of neighbors with shall-issue laws, and an interaction between the own and neighboring shall-issue variables as explanatory variables. The own effects of the law reduce violent crimes and increase property crimes, especially auto theft and larceny. The neighboring shall-issue coefficient measures the effect of a neighbor's law on a home county without a shall-issue law. Except for assaults, these spillover effects

³ To test for sensitivity, we tried two other definitions of neighbors: counties that were adjacent to each other or were within 100 miles of the geographic center of the home county. Neither definition altered the results shown here, though the spillovers were about 15-percent larger for counties within 100 miles than for those within 50 miles.

⁴ This is most important for murder and rape. See Lott and Mustard (1997) and Lott (1998a) for other approaches to deal with the problem of missing observations.

⁵ Trends are used for counties in states that enacted the law during the period studied. Prior to 1977 the last state to enact a shall-issue law was Washington in 1961.

TABLE 1—EFFECTS OF OWN AND NEIGHBORING SHALL-ISSUE LAWS ON CRIME RATES

Independent variable	Regression coefficients		
	Murder	Rape	Robbery
Part A:^a			
Own shall-issue dummy	-6.57 (3.848)	-4.15 (3.131)	-2.76 (1.908)
Neighboring shall-issue dummy	4.5 (1.676)	7.45 (3.59)	4.2 (1.851)
Own shall-issue × neighboring shall-issue	-8.7 (2.383)	-7.29 (2.580)	0.41 (0.131)
N:	25,931	33,121	34,271
Part B:			
Own shall-issue dummy	-1.401 [-16.1] (8.07)	-1.848 [-5.2] (3.80)	-15.00 [-6.8] (3.83)
Neighboring shall-issue dummy	0.816 [+9.4] (3.68)	-0.348 [-1.0] (0.56)	15.72 [+7.1] (3.14)
Own shall-issue × neighboring shall-issue	-0.815 [-9.4] (2.71)	0.321 [+0.9] (0.38)	-26.80 [-12.1] (3.94)
Difference in trends before and after law	-0.124 [-1.4] (5.22)	0.264 [+0.7] (3.97)	-4.81 [-2.2] (8.99)
Part C:			
Effect of own law in fourth year of law	-1.774 [-20.4] (8.37)	-1.011 [-2.8] (1.70)	-28.95 [-13.1] (6.05)
Effect of neighbor's law in fourth year, own law = 0	0.952 [+11.0] (2.45)	2.352 [+6.6] (2.16)	35.87 [+16.2] (4.09)
F test: own effect + neighbor's effect = 0	$p = 0.0076$ $p = 0.2940$ $p = 0.5011$		
Effect of neighbor's law in fourth year, own law = 1	0.140 [+1.6] (0.33)	2.713 [+7.6] (2.24)	9.34 [+4.2] (0.96)

Notes: The regressions in part A use the log of the crime rate and the arrest rate that corresponds directly to the crime rate being evaluated. Parts B and C were run on the actual crime rate, and the implied percentage change in crime rates as evaluated at the mean is reported in brackets. Because parts B and C use the violent-crime arrest rates for all the violent-crime categories and the property-crime arrest rates for all the property-crime categories, the sample sizes for those regressions are 44,445 for violent-crime categories and 42,326 for property-crime categories. The absolute values of t statistics are shown in parentheses.

^a Main table entries in part A are percentages.

estimated spillover effect beneficial, reducing neighboring assault rates by 3.6 percent.

Adding the coefficients for the neighboring shall-issue variable and the interaction effect provides the effect of a neighbor's law if the home county already has a shall-issue law. In crime categories where the neighboring effect is either economically important or even marginally significant (murder, rape, robbery, assault, property crime, and burglary) the interaction effect is always of the opposite sign. Therefore the magnitude of spillover effects are mitigated for counties that have already enacted shall-issue legislation.

Part B allows for spillover effects and trend-rate effects of the own-county law, as well as the changes described in Section I, using the aggregate arrest rates and the crime rate instead of the logarithm of the crime rate as the endogenous variable. Consider the estimates for the violent crime rate. The coefficient on the own shall-issue dummy is -34.16, implying a decrease in the violent crime rate of 34.16 per 100,000 residents, relative to what would have occurred in the absence of the law (a decrease of 5.9 percent). The trend coefficient is reduced by 3.46 after passage of the law, so that in the second and third years of the law the violent crime rate is reduced by 37.62 (34.16 + 3.46) and 41.08 (34.16 + [2 × 3.46]) per 100,000 residents, respectively. In every crime category except larceny, the short-run impact of the own shall-issue law significantly reduces crime rates. For murder, robbery, property crime, burglary, and auto theft, the deterrent effect of the law grows over time, while it gradually diminishes for assaults and rapes.⁶ Except larceny, the own effect of the law reduces all crimes over the first seven years.

The second row in part B indicates that spillover to neighbors tends to be deleterious: the murder rate, robbery rate, overall property crime rate, and burglary rate all significantly increase (only assaults decline). Some of these changes are quite large: murders increase by 9.4 percent, and robberies increase by 7.1

are either deleterious or insignificant. Some spillover effects are substantial: rapes increase by 7.45 percent, robberies by 4.2 percent, and murder by 4.5 percent. Only for assaults is the

⁶ In part B, if the dependent variables are instead the logs of the crime rates, the trend effects of the own law are significantly negative for all crime categories, including rape and assaults.

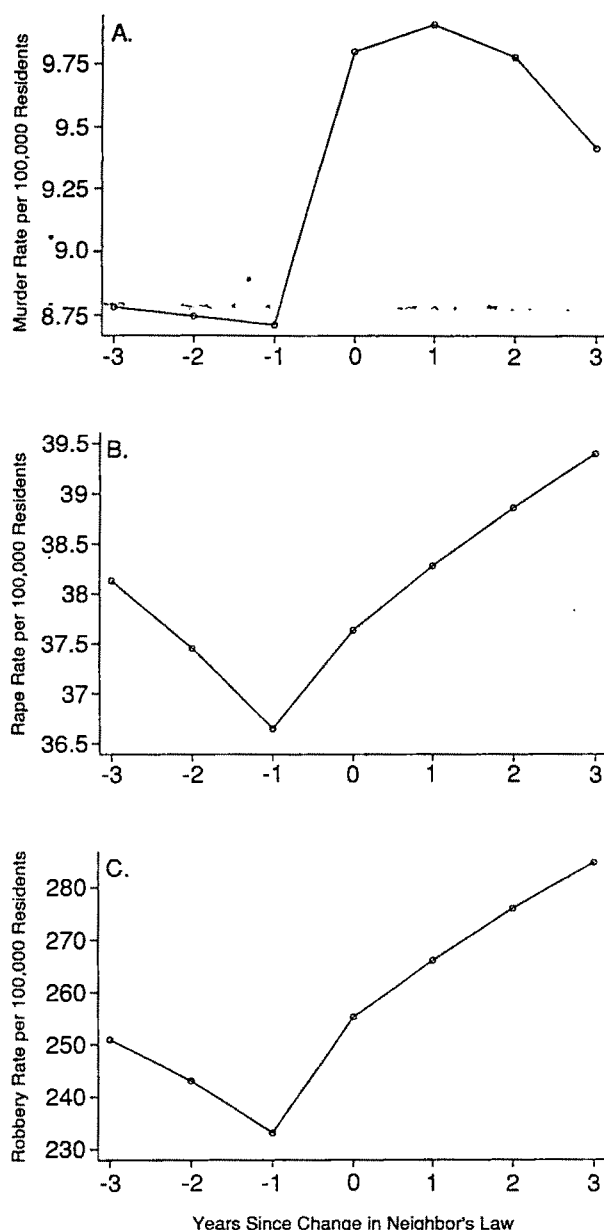


FIGURE 1. THE EFFECT OF A NEIGHBOR'S SHALL-ISSUE LAW ON CRIME: (A) EFFECT ON OWN MURDER RATE; (B) EFFECT ON OWN RAPE RATE; (C) EFFECT ON OWN ROBBERY RATE

percent in nearby counties without a shall-issue law.

If the home county already has a shall-issue law, the impact of the neighbor's law is the sum of the neighbor's law and interaction coefficients. In every crime category (other than larceny) the own-county \times neighboring-county interaction has the opposite sign of the neighboring-county effect. Thus the home county appears largely insulated from the del-

eterious effects of a neighbor's shall-issue law if the home county already allows citizens to carry concealed weapons. If the home county has a right-to-carry law, the only spillover effects that can be statistically determined from a neighbor's shall-issue law are beneficial: decreases in larceny and overall property-crime rates.

Parts A and B of Table 1 do not allow for any trend effects of the neighbor's shall-issue law on the home county. We now modify the specification in part B by including time-trend variables before and after the neighbor's right-to-carry legislation was enacted and present results from these regressions in part C of Table 1. Rather than reporting all the shall-issue variables, we summarize the results by only presenting estimates for the impact of shall-issue laws in their fourth year. We consider three types of changes in shall-issue laws: the own-county effect alone, the neighbor or spillover effect if the home county does not have a shall-issue law, and the spillover effect if the home county has a shall-issue law. Consider first the impact of a change in own-county right-to-carry laws. The own-law effects in part C are virtually identical to those in part B: allowing citizens to carry concealed weapons leads to significant reductions in crime rates in every crime category (other than larceny). Murder and robbery respectively decline by 20 percent and 13 percent.

The second row of part C measures spillovers to counties without a right-to-carry law. A shall-issue law in neighboring counties leads to significant increases in murders, rapes, robberies, property crimes, burglaries, and larcenies. Many of these deleterious spillover effects are quite large. For rapes, property crimes, burglary, and larceny, these spillover effects are even larger than the own shall-issue effect. The only evidence of beneficial spillover effects is for assaults and auto theft.

The final row of part C presents estimated spillover effects for a home county with the shall-issue law. In general, the presence of an own shall-issue law appears to mitigate the deleterious spillover effects from a neighbor's shall-issue law. In particular the harmful spillover effects for murder, robbery, property crimes, and burglary are eliminated by the presence of an own-county shall-issue law.

Harmful spillover effects remain for rapes and larceny, and there are beneficial spillover effects for larceny.

The spillover effects are vividly illustrated in Figure 1, which is based on part C of Table 1 and the inclusion of additional squared time trends for the periods before and after the adoption of the law. Year 0 is the year when a neighbor's shall-issue-law takes effect. All four of the violent-crime categories show dramatic crime spillovers precisely at year 0.

We also investigated whether changing arrest rates in neighboring areas similarly affect crime in the home county. Although we consistently find a strong negative relationship between the arrest rate and crime for the home county, we cannot detect any spillover effects for arrest rates. With one exception, the coefficients on neighboring arrest rates are at most one-eighth as large as the coefficient on the own arrest rate, and the *t* statistics are never significant.

III. Conclusion

The benefits that a county obtains from its state passing a shall-issue concealed-handgun law are generally stronger than those found in previous work. Spillover effects on neighboring areas are almost always deleterious. Criminals tend to move across communities more readily in response to changes in concealed-handgun laws than in response to changes in

arrest rates. The spillover effects are surprisingly large, especially for property crimes, thus challenging existing research which ignores these considerations. The spillovers are immediate and increase over time (with the exception of assaults and auto theft). Except for rapes, the negative effects of a neighbor's law are mitigated by adoption of the law by one's own state. Taken together, these results imply that concealed handguns deter criminals and that the largest reductions in violent crime will be obtained when all the states adopt these laws.

REFERENCES

- Bartley, William Alan and Cohen, Mark A. "The Effect of Concealed Weapon Laws: Estimating Model Uncertainty." *Economic Inquiry*, April 1998, 36(2), pp. 258-65.
- Daniel, Kermit and Lott, John R., Jr. "Should Criminal Penalties Include Third-Party Avoidance Costs?" *Journal of Legal Studies*, June 1995, 24(2), pp. 523-34.
- Lott, John R., Jr. "The Concealed Handgun Debate." *Journal of Legal Studies*, January 1998a, 27(1), pp. 221-43.
- . *More guns, less crime: Understanding crime and gun control laws*. Chicago: University of Chicago Press, 1998b.
- Lott, John R., Jr. and Mustard, David B. "Crime, Deterrence, and Right-to-Carry Concealed Handgun Laws." *Journal of Legal Studies*, January 1997, 26(1), pp. 1-68.

Engaging Students in Quantitative Analysis with Short Case Examples from the Academic and Popular Press

By WILLIAM E. BECKER *

"Students have unnecessary difficulty learning economics because textbooks generally do not have enough good examples of real-world applications."

—Gary Becker

(*Business Week*, 21 October 1996)

Examples, problems, and case studies that incorporate events reported in the newspapers, magazines, and journals that public administrators, business executives, and professors read can enhance the teaching of statistics and econometrics. I illustrate why and how students should be engaged in "headline" situations to motivate analyses.

I. How Economics Courses Are Taught

Following the completion of micro- and macroeconomics principles courses, for which the majority of students are enrolled to fulfill requirements for other majors, economics majors take two intermediate micro and macro courses, a course in statistics/econometrics, and some field courses that may or may not include more quantitative methods (John Siegfried et al., 1991). In a national survey, Michael Watts and I found some differences between the way statistics and econometrics

courses are taught and the way the other undergraduate economics courses are taught (Becker and Watts, 1996). In particular, problem sets are used more in statistics and econometrics than in other undergraduate economics courses. Curiously, however, those applications are not based on events reported in newspapers, magazines, and journals that economists read. How timely and relevant can problem sets be if they are not documented in current events?

Ideally instructors set problems raised by their own research and consulting; problems students can expect to see on their jobs. After all, the rationale for teaching statistics and econometrics outside a mathematics department rests on a belief that there is something special about economic analyses. That is, economists' use of statistics is tied to the issues they face. Although the calculation of a mean and a median, for example, is the same in medicine and economics, a discussion of the average duration of economic expansions since World War II is more pertinent to those majoring in economics than is a discussion of average blood pressure or average time to dementia with mad-cow disease. The importance of economic theory is often lost when mathematicians attempt to make situations real, as seen for example in the "Chance Course" (J. Laurie Snell and John Finn, 1992), where a potpourri of statistical applications are presented with no discipline-grounded analyses.

To teach students to apply the tools of statistics to actual situations and data encountered by economists, there is little justification for examples involving the drawing of balls from urns, flicking of spinners, tossing of coins, or contrived card and dice tricks. Yet these methods of generating data continue to be found in the activity-based teaching and assessment

[†] *Discussants:* William Greene, New York University; Robin Lumsdaine, Brown University; Kim Sosin, University of Nebraska-Omaha.

* Department of Economics, Indiana University, Bloomington, IN 47405. This work is supported by grants from the National Science Foundation (DUE 955408 and DUE 9653421) and a grant from the Calvin Kazanjian Economics Foundation. Parts of this manuscript were prepared at the University of South Australia where I was an adjunct professor, in residence for the period of May–August 1997. Suzanne Becker, Peter Kennedy, and George Bredon provided constructive criticism on earlier drafts.

methods advocated by mathematics educators (Richard Scheaffer, 1996; Iddo Gal and Joan Garfield, 1997). Students need to be involved in working exercises, considering case studies, and solving problems that reflect what economists do in their research. This implies replacing the urns with preselection pools from which individuals are hired, replacing the coins with surveys in which individuals face multiple-choice responses, and replacing examples from genetics with examples from quality control. If games of chance are employed to demonstrate risk, then use actual situations involving lotteries or other financial decisions students see in the popular press.

Basic and more advanced quantitative methods can be made "real" with reference to the popular press. But simply assigning students to find applications and relevant data in newspaper articles is too unstructured as a starting point. As demonstrated in Becker (1997), the mini-case study approach introduced by Rendigs Fels (1974) in the teaching of economics can be modified to lead students into analysis. Short case studies enable instructors to demonstrate a specific form of analysis while giving students an opportunity to observe how concepts and theories are used to examine a diverse array of issues. With short and focused case studies students do not get lost in extraneous verbiage as they begin a study of statistics and econometrics.

Few economists have published or given courtroom or boardroom testimony in all the areas found in statistics and econometrics courses. The popular press and academic journals can be used as a surrogate for missing credentials and experiences, although we may feel more comfortable and find it easier to work with made-up problems and fake data. The discrepancy between fairy tales and fact, however, are apparent to students. Consider, for example, two questions taken from different end-of-semester examinations given in the 1996 and 1997 spring semesters at Indiana University.

Q 1996: Economic theory suggests that the greater the number of pages in a book, the greater its price. As a novice economist, you want to test for a significant relationship be-

tween price (y , in dollars) and number of pages (x). You run a regression and obtain

$$\hat{y} = 1.041553 + 0.009907x.$$

If a book contains 900 pages, the predicted price of that book would be

- A. \$8.92
- B. \$9.96
- C. \$675.24
- D. \$936.01.

Q 1997: A real-estate report in the *Wall Street Journal* (28 February 1997) gave "the top 10 condominium sales in New York City." The estimated relationship between condo sale price (y , in millions of dollars) and condominium size (x , in square feet) is given below. Based on this regression, was the condominium at 353 Central Park West ($y = 1.6$, $x = 2,733$) a good buy?

$$\hat{y} = 0.8456 + 0.001085x$$

- A. No, because its actual price (\$1.6 million) was less than its predicted price (\$2.12 million).
- B. No, because its residual ($y - \hat{y}$) is negative; a good buy has a large positive residual.
- C. Yes, because its actual price (\$1.6 million) was less than its predicted price (\$2.12 million).
- D. Yes, because its sale price per square foot is relatively low at \$585.44 per square foot.

Both exams had over 40 multiple-choice questions with similar averages of about 65-percent correct. As represented by Q1997, the use of cited sources gives a sense of realism to the 1997 exam. Questions such as Q1996 provide no references and are obviously contrived: a 900-page book priced at \$9.96 must seem ludicrous to students who buy textbooks.

To the extent that exam questions reflect the manner in which a course is taught, I suspect that students who took the course represented by Q1997 walked away with better statistical skills and different impressions of their applicability than those completing the 1996 course. This of course is a testable hypothesis,

which I regret has not been explored. Nevertheless, the Q1997-type of question makes clear that the course is aimed at applying statistics to economics.

Although not usually viewed as such, test questions can be used to stimulate student interaction. For example, 12 times during the semester, typically as the course moves between topics, I project a single multiple-choice question on the screen of a 350-student lecture hall. Each student answers the question by marking A–E on a machine-scorable sheet that is distributed as they enter the hall. For a second question, “A. Certain” to “E. Doubtful” is marked as an expression of confidence. Students then discuss their answers with neighbors; the lecture hall is abuzz. After a few minutes, students repeat the process with question 3 corresponding to question 1 and final confidence being question 4. Student attendance and participation increased with the introduction of this activity because students get credit for attempting the four responses, as well as for selecting the correct response to question 3. I call this innovation “a class-participation quiz”; its origins can be traced to Eric Mazur, a Harvard University physics professor.

A belief likely held by all is that quantitative methods courses should contribute more to student development than what is typically measured by paper-and-pencil tests. I suspect that many also believe that there is a difference between the type of response evoked from multiple-choice (fixed-response) questions and essay or short-answer (construct-response) questions. But in what way do these different forms of time-constrained testing capture different dimensions of student performance? The debates about the merits of multiple-choice versus essay exams may be missing the broader spectrum of educational outcomes students are to achieve and the way in which they are achieved. As educators, we need to move beyond simple measures of knowledge to consider what leads to student persistence during a course, into another course, or into a major, and what skills students need for future performance in the workplace. The need for communication skills and interaction with others is made clear in the Harvard Assessment Seminars survey

(Richard Light, 1992 p. 8). My participation quizzes are an attempt to get students to interact in a large class setting. I mention this activity only to show that quizzes and tests can be used to advance communication skills and student interaction as they work to acquire quantitative skills even in large lecture halls where such activity is usually considered inefficient if not impossible.

II. What Is Taught

The learning of statistics often starts with the calculations of descriptive statistics. The amount of experience students should get with computational formulas is debatable, but sooner or later they must be exposed to computer packages. Although writing equations on a blackboard is traditional, teaching statistics and econometrics without spending time in a computer lab is difficult to justify. Data from the press can be employed even for the most basic calculations. Small data sets enable hand calculations to verify what a computer program does. One such data set is the duration of economic expansion. A *Wall Street Journal* (15 November 1996 p. A1) article stated, “The nation’s continuing economic expansion, now 67 months old, has far surpassed the previous postwar average (50 months), and few see the party ending anytime soon” and provided the monthly durations of the 10 expansions that started since World War II. Students can be engaged in the analysis of these data with questions that require them to write short answers in class that can be discussed or graded as in a one-minute paper activity. They quickly realize the consequence of censoring and truncation as they address the calculation and interpretation of the median and mean given that the March 1991 economic expansion had not ended as of November 1996.

While on the topic of the economy, a multiple-part feature in the *Washington Post* (15 October 1996 p. A1) reported on the public’s versus economists’ views: “Good news for President Clinton: There is a core group of voters who agree with him that the economy is stronger than it was when he took office ... Bad news for the president: They’re econo-

mists—and nobody believes them anyway.” Our low credibility likely affects our ability to teach data analysis. Teaching econometrics was not helped by Marilyn vos Savant’s answer (*Parade Magazine*, 5 September 1997 p. 22) to a reader who asked how economists working with the same data reach different conclusions. She wrote that economists are like chefs: one is amazed at the variety of stuff they cook up when given exactly the same ingredients, equipment, and staff. But as students need to learn, it is often the sample data with which economists work that gives rise to the differences in interpretation. High-school science classes as well as the news media do a good job of indoctrinating students in the need for “good scientific methods,” but neither the science teacher nor the newscaster typically sheds light on problems in sampling and estimation.

Consider the “margin of error” for survey results reported by the news media. Readers and television audiences must be expected to know what this information is, because the journalist never elaborates. Curiously the survey margin-of-error calculation is not in most introductory statistics textbooks. In the introductory statistics course, I use reported survey statistics to show the calculation of the “margin of error,” $\pm 1/\sqrt{n}$. The discussion emphasizes the critical nature of two assumptions: (i) that there is a 50-percent chance that the respondent will select one alternative versus another on each of the questions in the survey and (ii) that the estimator is normally distributed so that a 95-percent confidence interval is approximately plus or minus two standard errors of an estimator. I emphasize that these are technical details but the more interesting questions are related to sampling methods, and the reliability and validity of the measurement instruments. Students, however, appear fascinated with the simplicity of the $\pm 1/\sqrt{n}$ calculation and Dan Rather’s authoritative voice when he says “margin of error.” In the more advanced econometrics class, I emphasize the problems of selection and nonresponse error. As in Morris Hansen and William Hurwitz (1946), students come to realize that the size of the sample alone may be a poor indicator of reliability. From Jessica Utts (1991) they see that replication and statistical

significance are not the simple ideas portrayed in textbooks.

Some instructors attempt to make the randomness in sampling real by having students do surveys. The use of experimental economics and games has also become popular in the teaching of economics. I question whether students see the reality in what they are doing unless there is a visible cue to events outside the educational arena. Self-reported data collected in classroom surveys or in the community are typically more appropriately used to show unreliable and invalid collection methods and opportunistic samples than a good source of data with which to work in a statistics course. To tie to the “real world,” these classroom data-collection problems can be likened to the efforts to establish the employment effect of the 1992 change in the New Jersey minimum-wage law, as featured in numerous headlines. The debate over the use of wage data collected in phone interviews of restaurant workers in Pennsylvania and New Jersey versus payroll data supplied by a trade association is evident in the exchange of viewpoints and reader reactions found in *Business Week* (24 April 1995 and 15 May 1995) and Gary Becker’s column in *Business Week*, “How Bad Studies Get Turned Into Bad Policies” (26 June 1995).

In introductory statistics students learn about probability distributions, but these are seldom tied to real-world applications. A way I solve this problem can be seen in a binomial-distribution application involving an *Insider TV* show (15 February 1995) that showed a tow-truck operator stealing items from one of four cars the television show placed in a New York City tow area. A New York City official is quoted as saying that no more than 1 percent of the cars that are towed and impounded have objects inside stolen. In a “think, pair, share activity,” I ask students to assess the manner in which a one-in-four theft rate can be considered unusual if one in a hundred is expected. They come up with the question “How far above 0.04 is 1.00?” If the calculation is not forthcoming, I ask them for the probability of at least one car in four being burglarized if the probability of a burglary is 0.01. The appropriateness of the binomial model and the magnitude of the resulting 0.039 probability

are discussed. The debate over the television show's conclusion versus the implications of this 0.039 probability is not dull and quickly leads to a discussion of the difference between statistical significance and importance when working with small versus large samples.

The popular press is riddled with errors in data, calculations, and statistical reasoning, as demonstrated recently by Cynthia Crossen (1994). Everyone has favorite examples, especially when it comes to misleading graphical presentations. But even here the press can be used to give credibility to proper graphical techniques. For example, *Fortune* (27 October 1997) ran a feature article on Yale professor Edward Tufte's (1983) book on displaying data, calling it the guide for those who work with numbers. Could there be a greater endorsement for proper data display?

Crossen (1994) and Tufte (1983) demonstrate that the popular press is loaded with statistical nonsense. Using errors found in the press for classroom demonstrations and discussions, however, may be risky. Journalists may be viewed as more credible than academics, as suggested by the previously cited *Washington Post* quote: "They're economists—and nobody believes them." A teacher's criticism, no matter how legitimate, may fall on deaf ears. Furthermore, even if the teacher is successful in demonstrating the error, students may view this as more evidence of the irrelevance of econometrics; after all, the journalist draws a large salary writing for a big-name press and apparently did not need to know statistics or econometrics to get his or her job! Furthermore, there is always the chance that the criticism could be erroneous. Recall all of those Ph.D.'s who wrote to Marilyn vos Savant at *Parade Magazine* claiming she was "the goat" in Monty's three-door game-show problem. *The New York Times* gave front-page coverage to this "Let's Make a Deal" controversy. J. P. Morgan et al. (1991), John Georges and Timothy Craire (1995), and other probabilists and game theorists finally had to admit that Marilyn was right.

One way to avoid errors is to restrict use of the popular press to articles based on the work of cited researchers. This approach makes use of reporters' ability to identify newsworthy

work and the researchers' expertise as statisticians/econometricians. It may also yield the instructor great data sets. For example, *The New York Times* (11 April 1994) reported on how Federal District Court Judge Clarence Newcomer turned to Orley Ashenfelter to help him decide whether to order a new election in Pennsylvania's Second State Senatorial District in Philadelphia or declare the losing candidate Bruce Marks, a Republican, the winner in the November 1993 election. In this special election to fill a Senate vacancy, Marks received 19,691 votes on voting machines and his Democrat opponent, William Stinson, received 19,127. But in absentee ballots Stinson received 1,391 to Marks' 461. The Republicans charged that many of the absentee ballots were falsified by the Democrat-controlled County Board of Election. The scatter plot of absentee and machine-ballot differences in the previous 21 elections that accompanied the *New York Times* article makes clear the role of regression analysis. An advantage of this data set for classroom use is that it is small and yet provides a meaningful estimation of a simple regression model.

Ashenfelter's work with small data sets often appears in the popular press and refereed scholarly journals as well. It is thus ideally suited for classroom use. As another example, consider his study with Alan Krueger on the return to schooling (Ashenfelter and Krueger, 1994). This study was highly publicized in the popular press both for its use in U.S. Congressional debate and in discussions of Richard Herrnstein and Charles Murray's (1994) controversial book on IQ, as seen for example in *The New York Times* (9 November 1994). The fixed-effect, least-squares estimates for the difference in identical twins' earnings as a function of differences in years of schooling involve only 149 observations and like the Pennsylvania election data provide another example of a meaningful simple regression. Unlike the Pennsylvania election data, however, this data set can be used to demonstrate problems caused by selection effects and measurement error, with generalized least squares used to correct these problems.

The press is rich in articles showing sample-selection problems. For example, William Greene (1997) introduces the effect of trun-

cation with a story from the *New York Post* on "affluent Americans" having average income of \$142,000 per year. He shows how the income of the "typical American" can be estimated from the information provided about the top 2 percent who make at least \$100,000 per year (p. 683). Similar examples of truncation appear regularly in the press. One of my favorites has the headline "Sears Is Accused of Billing Fraud at Auto Centers." This *Wall Street Journal* (12 June 1992) article reported a year-long study by the California Department of Consumer Affairs in which "its agents were overcharged nearly 90% of the time, by an average of \$223." The question: what happened the other 10 percent of the time?

III. Engaging Students

To engage students individually and in group work with computers, small classes are needed. In the Business and Economics Statistics class at Indiana University approximately 800 students per semester are allocated to some 32 computer-lab sections. Computer lab instructors have weekly staff meetings to review what students are to do, and to learn what problems to expect. Two lab activities that students seem to like require them to go to the World Wide Web in search of information. For the first, students are to find a news item for which a multiple-choice question can be written. Students are encouraged to use the Web and a search engine (Yahoo) to find the major news and data sources. Although all students attempt this, they quickly learn that the "World Wide Wait" is frustrating. Most assignments are handed in with an original hard-copy source and not a printout from the Web. In the honors section of the course, students must find a data source that can be used throughout the semester for data analysis. At the beginning of the semester students are told to find data on several variables for which they suspect a relationship. Students are encouraged to search the Web. Students learn that all the major government agencies and news services maintain Web sites where data can be accessed. They also learn that the most comprehensive listing of these sites is at:

<http://econwpa.wustl.edu/EconFAQ/EconFAQ.html>

where Bill Goffe's opening page has over 44 links to resources for economics. For current information, students no longer need to visit the reading room or government-documents section of the library. Statistics and econometrics courses can be made timely and relevant right in the computer labs.

REFERENCES

- Ashenfelter, Orley and Krueger, Alan. "Estimates of the Economic Return to Schooling from a New Sample of Twins." *American Economic Review*, December 1994, 84(5), pp. 1157-73.
- Becker, William. *Statistics for business and economics using Microsoft Excel 97*. Bloomington, IN: SRB Publishing, 1997.
- Becker, William and Watts, Michael. "Chalk and Talk: A National Survey on Teaching Undergraduate Economics." *American Economic Review*, May 1996 (*Papers and Proceedings*), 84(2), pp. 448-53.
- Crossen, Cynthia. *Tainted truth: The manipulation of facts*. New York: Simon and Schuster, 1994.
- Fels, Rendigs. "Developing Independent Problem-Solving Skills in Economics." *American Economic Review*, May 1974 (*Papers and Proceedings*), 64(2), pp. 403-7.
- Gal, Iddo and Garfield, Joan, eds. *Assessment of challenge in statistics education*. Amsterdam: IOS Press, 1997.
- Georges, John and Craire, Timothy. "Generalizing Monty's Dilemma." *Quantum*, March/April 1995, pp. 17-21.
- Greene, William. *Econometric analysis*, 3rd Ed. Englewood Cliffs, NJ: Prentice Hall, 1997.
- Hansen, Morris and Hurwitz, William. "The Problem of Non-response in Sample Surveys." *Journal of the American Statistical Association*, December 1946, 41(236), pp. 517-29.
- Herrnstein, Richard and Murray, Charles. *The bell curve*. New York: Free Press, 1994.
- Light, Richard. *The Harvard Assessment Seminars*, 2nd Report. Cambridge, MA:

- Harvard University Graduate School of Education and Kennedy School of Government, 1992.
- Morgan, J. P.; Chaganty, N. R.; Dahiya, R. C. and Doviak, M. J. "Let's Make a Deal." *American Statistician*, November 1991, 45(4), pp. 284-87.
- Scheaffer, Richard L. *Activity-based statistics*. New York: Springer-Verlag, 1996.
- Siegfried, John; Bartlett, Robin L.; Hansen, W. Lee; Kelley, Allen C.; McCloskey, Donald N. and Tietenberg, Thomas H. "The Status and Prospects of the Economics Major." *Journal of Economic Education*, Summer 1991, 22(3), pp. 197-224.
- Snell, J. Laurie and Finn, John. "A Course Called 'Chance.'" *Chance*, 1992, 5(3-4), pp. 12-16.
- Tufte, Edward. *The visual display of quantitative information*. Cheshire, CT: Graphics Press, 1983.
- Utts, Jessica. "Replication and Meta-analysis." *Statistical Science*, November 1991, 6(4), pp. 363-78.

Teaching Undergraduate Econometrics: A Suggestion for Fundamental Change

By PETER E. KENNEDY *

Contrary to the belief of most econometrics instructors, upon completion of introductory statistics courses the vast majority of students do not understand the basic logic of classical statistics as captured in the sampling-distribution concept. They have learned to do mechanical things such as compute a sample variance, run a regression, and test an hypothesis, and they know they can pass the course by remembering how these techniques work. They view statistics as a branch of mathematics because it uses mathematical formulas, so they look at statistics through a mathematical lens.

What they are missing is the statistical lens through which to view this world, allowing this world to make sense. The sampling-distribution concept is this statistical lens. My own experience discovering this lens was a revelation, akin to the experience I had when I put on my first pair of eyeglasses: suddenly everything was sharp and clear. In this paper I discuss this issue and suggest a means whereby instructors of undergraduate econometrics courses (i.e., following introductory statistics) can provide the missing statistical lens.

I. The Importance of the Sampling-Distribution Concept

"Constructivism," a recent theory of learning, has been widely accepted in education communities and is the guiding theory for much research and reform in mathematics and science education. According to this theory, students bring to the classroom their own ideas, and rather than passively adding to these ideas as material is presented in class, they actively restructure the new information to fit it into their own cognitive frameworks. In this way they are "constructing" their own knowl-

edge, rather than copying knowledge delivered to them through some teaching mechanism. Joan Garfield (1995 p. 30) provides this explanation:

Regardless of how clearly a teacher or book tells them something, students will understand the material only after they have constructed their own meaning for what they are learning. Moreover, ignoring, dismissing, or merely "disproving" the students' current ideas will leave them intact—and they will outlast the thin veneer of course content.

Students do not come to class as "blank slates" or "empty vessels" waiting to be filled, but instead approach learning activities with significant prior knowledge. In learning something new, they interpret the new information in terms of the knowledge they already have, constructing their own meanings by connecting the new information to what they already believe. Students tend to accept new ideas only when their old ideas do not work, or are shown to be inefficient for purposes they think are important.

An important consequence of this theory is that students whose personal cognitive framework does not match that of the instructor/textbook find it difficult to learn what the instructor wants them to learn, because what is taught in the classroom appears to be a set of unrelated ideas that must be memorized. The result is frustration, poor understanding of course material, and a dislike of the subject. I believe that this problem characterizes econometrics, and that the main culprit is that students come to econometrics, at both the undergraduate and graduate level, not understanding adequately the sampling-distribution concept.

As illustrated below, this concept provides a unifying logic that permits estimation,

* Department of Economics, Simon Fraser University, Burnaby, BC, Canada V5A 1S6.

hypothesis testing, and econometric's algebraic manipulations to be seen in context:

- (i) Using a formula β^* to produce an estimate of β can be conceptualized as the econometrician shutting his or her eyes and obtaining an estimate of β by reaching blindly into the sampling distribution of β^* to obtain a single number.
- (ii) Because of (i) above, choosing between β^* and a competing formula β^{**} comes down to the following: Would you prefer to produce your estimate of β by reaching blindly into the sampling distribution of β^* or by reaching blindly into the sampling distribution of β^{**} ?
- (iii) Because of (ii) above, desirable properties of an estimator β^* are defined in terms of its sampling distribution. For example, β^* is unbiased if the mean of its sampling distribution equals the number β being estimated.
- (iv) Because of (iii) above, econometricians spend a lot of algebraic energy figuring out sampling-distribution properties, such as mean and variance.
- (v) The properties of the sampling distribution of an estimator β^* depend on the process generating the data, so an estimator can be a good one in one context but a bad one in another; β^* 's sampling-distribution properties need to be recalculated for every different data-generating process.
- (vi) All statistics, not just parameter estimates, have sampling distributions. For example, under the null hypothesis, an F statistic will have a sampling distribution described by the F table found in statistics textbooks.

II. Do Students Understand the Sampling Distribution Concept?

There is much research, summarized for example by Garfield and Andrew Ahlgren (1988), showing that a large proportion of university students in introductory statistics courses do not understand many of the concepts they are studying. But I cannot refer to this research to defend my claim that students do not understand the sampling-distribution

concept, because this research relates mainly to the psychology of probability in which students can be shown to use misleading heuristics to compute probabilities. Instead I defend my claim as follows:

- (i) I have sampled scores of students at the beginning graduate and upper-level undergraduate level, finding that very few are able to explain what is a sampling distribution. Most think it is a distribution pictured by drawing a histogram of the sample data.
- (ii) I have delivered to new graduate students an expository lecture on the sampling-distribution concept and the role it plays in statistics/econometrics, and through anonymous surveys after this lecture have discovered that almost all students admit that they did not realize the important conceptual role of the sampling distribution in econometrics.
- (iii) I frequently encounter students with A grades in their introductory statistics courses who clearly have no understanding of statistics beyond a mechanical ability to apply standard statistical procedures.
- (iv) I often see graduate students with an impressive ability to derive statistical formulas but a remarkable inability to explain how to evaluate those formulas via a Monte Carlo study, even after Monte Carlo procedures have been explained.
- (v) Others have expressed a similar sentiment. Ruth Hubbard (1997 p. 11) complains that "When graduate students from a variety of disciplines approach me for help with analysing or interpreting their data, I always begin by asking if they understand the statistical term 'standard deviation.' Their standard response is, 'It is something that you calculate from a formula.' In most cases no amount of probing succeeds in clarifying either the formula or what it measures." Vijaya Duggal (1987 p. 26) contends "that more than three-fourths of students finishing a course in quantitative methods who know the mechanics of hypothesis testing have no conceptual

comprehension of the sampling distribution of the mean.”

How has this state of affairs come about? It is not because the introductory statistics books ignore sampling distributions; they all have plenty of good material on this concept and give it appropriate emphasis. And it is not because instructors ignore this dimension of introductory textbooks; all instructors swear that they teach this concept thoroughly. I suggest three reasons why students fail properly to understand this concept. First, it is difficult. Students can visualize a distribution of sample observations, but a sampling distribution is at a higher level of abstraction, where sample observations yield a single value of a statistic, not an entire distribution. Second, texts typically introduce the sampling-distribution concept in the context of the sample mean statistic but do not follow it up adequately when discussing regression, the central focus of econometrics. Third, and I believe most important, as Hubbard (1997 p. 1) quotes Lauren Resnick “We get what we assess, and if we don’t assess it, we won’t get it.” Students are utility-maximizers. Typically their exams require an ability to interpret regressions, calculate t statistics, and perform hypothesis tests, but not an ability to explain the concept of a sampling distribution. They cannot imagine, and instructors and textbooks seldom provide, examples of exam questions that in any meaningful way probe student understanding of this concept.

Good instructors realize that students have mostly forgotten their introductory statistics material, or never learned it properly in the first place, and proceed on the basis that this introductory material needs to be reviewed. Unfortunately, this review suffers from the same problem noted earlier: there is little motivation for students to learn properly the concept of a sampling distribution and the role it plays in econometrics because instructors’ expositions of this concept, however clear, are seldom accompanied by appropriate example exam questions. Such questions provide motivation and, more importantly, force students to work out answers. Although lecturers like to think otherwise, brilliant expositions seldom cause students fully to understand; such

understanding comes through working out problems based on the concept to be learned. An old Chinese proverb bears repeating: I hear, I forget; I see, I remember; I do, I understand.

Econometrics textbooks do not help much. They are in too big a hurry to produce the theorems, proofs, and formulas that define theoretical econometrics. Some review introductory statistics but do so without much emphasis on sampling distributions and seldom provide meaningful questions that force student understanding of the sampling-distribution concept. I am amazed that econometrics textbooks pay so little attention to the sampling-distribution concept, but it is consistent with my observation that most econometrics instructors believe that their students already have a good understanding of this concept. This impression is reinforced by reading the very limited literature on teaching econometrics. Eric Sowe (1983) and his commentators, for example, ignore this issue. This is indeed a major part of the problem. If instructors are not aware of this deficiency in student understanding, they will not try to remedy it.

III. A Suggestion for Change

Two changes are needed. First, instructors need to produce for students a good exposition of what is a sampling distribution, couched in the context of regression coefficient estimation, followed by a discussion of the overall role of the sampling-distribution concept in econometrics. I have provided my own such exposition elsewhere (Kennedy, 1988a, b appendix A). An important ingredient of this exposition is an explanation of how, for a given data-generating context, econometricians learn the properties of statistics’ sampling distributions. This involves discussion of the roles of mathematical derivations and asymptotic algebra and, most importantly, an exposition of Monte Carlo studies: how computer simulation can be used to investigate the properties of a sampling distribution.

Second, a way must be found to hammer home this lesson so that the perspective it provides for later study is not forgotten. I suggest the use of “explain how to do a Monte Carlo

study" problems. By forcing students to outline step-by-step a Monte Carlo study to examine an econometric issue, they are forced to spell out very clearly their understanding of this issue. For example, I ask students to explain how to conduct a Monte Carlo study to examine the implications of omitting a relevant explanatory variable. Instructors may wish to ease students into this kind of question by providing the Monte Carlo instructions and asking students to anticipate the results, as in the following example:

- (a) Draw 50 x values from a uniform distribution between 3 and 12.
- (b) Draw 50 z values from a standard normal distribution.
- (c) Compute 50 w values as $4 - 3x + 9z$.
- (d) Draw 50 e values from a standard normal distribution.
- (e) Compute 50 y values as $2 + 3x + 4w + 5e$.
- (f) Regress y on x and save the x slope coefficient estimate $b1$.
- (g) Regress y on x and w and save the x slope coefficient estimate $bb1$.
- (h) Repeat this procedure from (d) to get 1,000 b and bb values.
- (i) Compute the averages of these sets of 1,000 values to get B and BB , respectively.
- (j) Compute the variances of these sets of 1,000 values to get VB and VBB , respectively.

Should B or BB be closer to 3? Should VB or VBB be closer to 0?

Kennedy (1998b appendix D) contains additional examples. Based on many years of experience using this approach I suggest several dos and don'ts (discussed at greater length in Kennedy [1998a]) for ensuring its success:

- (i) Only ask students actually to do a Monte Carlo study if you are willing to pay the high opportunity cost of having them learn how to program.
- (ii) Be prepared for the question: Won't our choice of parameter values affect the results of the Monte Carlo study?
- (iii) Be very careful evaluating students'

Monte Carlo descriptions because small mistakes, and especially vagueness, can reflect major misunderstandings.

- (iv) Begin with very easy questions such as finding the variance of the sample mean statistic.
- (v) Do not be surprised that, once the mechanical steps of a Monte Carlo study are in hand, students' biggest problem is how to simulate the data-generating process. If they cannot explain how to generate raw data for a heteroscedastic-error model, for a two-equation simultaneous-equation system, or for a qualitative-choice model, can you be confident that they understand what any of your algebra is about?
- (vi) In-class development of a step-by-step set of Monte Carlo instructions can be an effective means of generating classroom participation. Call on a student to suggest the first step, then go around the room asking each student in turn to supply the next step, or correct an incorrect step suggested by an earlier student.

As should be evident from the discussion above, this proposal requires that instructors spend considerable time on sampling distributions and Monte Carlo studies, something most do not now do. What should be given up to accommodate this new feature? I believe that the benefit of this innovation is so large that no instructor could possibly claim that the least-valued subset of their course has more benefit. Indeed, this innovation should enhance student understanding of all dimensions of a course, creating benefits beyond those attached solely to learning about data generation, sampling distributions, and Monte Carlo studies. Advanced estimation or testing techniques do not mean much to students if they do not understand the fundamental principles that lie behind them.

In my own courses, I have given up most mathematical derivations. Do we really want our students at the end of their course to be able to do technical things like derive the ordinary least-squares estimator and prove that it is BLUE? Such things have little meaning if the concept of a sampling distribution is not thoroughly understood.

IV. Conclusion

This paper urges fundamental change in the way in which undergraduate econometrics is taught. It argues that, at the beginning of the traditional undergraduate econometrics course (i.e., the course following introductory statistics), instructors should exposit carefully the concept of a sampling distribution, explain how Monte Carlo studies can be used to characterize sampling distributions, and, most important, follow this up throughout the course with assignments asking students to explain how to conduct Monte Carlo studies to investigate specific econometric questions.

Students find this approach intellectually very challenging because it forces them to think hard to understand an abstract concept. Julian Simon and Peter Bruce (1991 p. 29) make the same point in the context of their resampling approach to learning basic statistics, claiming that it "requires only hard, clear thinking. You cannot beg off by saying 'I have no brain for math!'" Some students do not rise to this challenge; they are more comfortable learning by rote a bunch of techniques and mathematical proofs and so find this approach more difficult.

An obvious corollary to the message of this paper should be emphasized. Instructors of beginning-level graduate econometrics courses need to realize that their students do not have a good understanding of the sampling-distribution concept and how it fits into econometrics. A story told by one of my former undergraduate students illustrates this. At the beginning of his graduate econometrics course he was struggling because he had to learn from scratch formulas and algebraic derivations that other students had encountered in their more traditional undergraduate econometrics courses. But he found that after the first month or so he had become

comfortable with the math and suddenly leapfrogged far ahead of the rest of the class because, he said, he understood what was going on but they understood only the mathematics.

REFERENCES

- Duggal, Vijaya G. "Coping with the Diversity of Student Aptitudes and Interests." *American Economic Review*, May 1987 (*Papers and Proceedings*), 77(2), pp. 24-28.
- Garfield, Joan. "How Students Learn Statistics." *International Statistical Review*, April 1995, 63(1), pp. 25-34.
- Garfield, Joan and Ahlgren, Andrew. "Difficulties in Learning Basic Concepts in Probability and Statistics: Implications for Research." *Journal for Research in Mathematics Education*, January 1988, 19(1), pp. 44-63.
- Hubbard, Ruth. "Assessment and the Process of Learning Statistics." *Journal of Statistics Education*, March 1997, 5(1), (www.stat.ncsu.edu/info/jse/v5n1/hubbard.html).
- Kennedy, Peter E. "Using Monte Carlo Studies for Teaching Econometrics," in W. E. Becker and M. Watts, eds., *Teaching undergraduate economics: Alternatives to chalk and talk*. Aldershot, U.K.: Edward Elgar, 1998a (forthcoming).
- . *A guide to econometrics*, 4th Ed. Cambridge, MA: MIT Press, 1998b.
- Simon, Julian L. and Bruce, Peter C. "Resampling: A Tool for Everyday Statistical Work." *Chance*, Winter 1991, 4(1), pp. 22-32.
- Sowey, Eric. "University Teaching of Econometrics: A Personal View," with comments by Jacques Dreze, Pascal Mazodier, Peter Phillips, Takamitsu Sawa, and Arnold Zellner. *Econometric Reviews*, 1983, 2(2), pp. 255-333.

AMERICAN ECONOMIC ASSOCIATION

PROCEEDINGS
OF THE
HUNDRED AND TENTH
ANNUAL
MEETING

CHICAGO, IL
JANUARY 3–5, 1998

KEVIN MURPHY

JOHN BATES CLARK MEDALIST

1997

Kevin Murphy is a brilliant economist whose skills span the full range of the discipline. He is a superb data analyst and econometrician who has provided a definitive description of fundamental labor-market phenomena such as wage differentials and patterns of unemployment. While his research is motivated by concrete economic problems and directed at understanding empirical issues, he is a gifted and original theorist. His work on economic growth reveals a penetrating intuition and a deep understanding of the mathematical structure of economics. He has been awarded the John Bates Clark Medal in recognition of his wide-ranging contributions to economics.

Wage inequality has received as much attention as any topic in labor economics during the past decade. Murphy's contributions to this field, reported in 18 published articles, are thorough, revealing, and pathbreaking. He was among the first to document the rise in wage differences among groups, to attribute the change to a shift in the demand for skill, and to show that inequality had also increased within groups. His studies of wage inequality are complemented by important work on patterns of unemployment.

Murphy's contributions to the study of wage inequality began with an exhaustive description of the radical changes in relative wage structures that took place during the early 1980's. ("The Structure of Wages," with F. Welch, 1992, in the *Quarterly Journal of Economics*). Murphy adopted a simple, but extraordinarily revealing, metric for describing changes in wage inequality through trends in wage quantiles ("Wage Inequality and the Rise in the Return to Skill," with C. Juhn and B. Pierce, 1993, in the *Journal of Political Economy*). He provided illuminating calculations of plausible ranges for changes in the time path of the demand for skill ("Changes in Relative Wages 1963–1987," with L. Katz, 1992, in the *Quarterly Journal of Economics*).

Murphy documented the fact that widening wage differences among groups were accompanied by increasing inequality within groups ("Wage Inequality and the Rise in the Returns to Skill," with C. Juhn and B. Pierce, 1993, in the *Journal of Political Economy*). He originated full-sample distribution accounting, a method for decomposing the sources of change in inequality and successfully applied this technique to show that the slowdown in convergence of black–white wage differentials reflected changes in the demand for skill ("Accounting for the Slowdown in Black–White Wage Convergence," with C. Juhn and B. Pierce, in M. Kosters, ed., *Workers and Their Wages*, 1991).

Murphy provided the facts with which theories of unemployment must contend ("The Evolution of Unemployment in the United States: 1968–1985," with R. Topel, 1987, in the *NBER Macroeconomics Annual*). He showed that the increase in U.S. joblessness from the 1960's to the 1980's was concentrated among the less skilled, paralleled the decline in wages of these workers, and was accompanied by a decline in labor-force participation ("Why Has the Natural Rate of Unemployment Increased Through Time," with C. Juhn and R. Topel, 1991, *Brookings Papers on Economic Activity*).

While Murphy's research on wage inequality and unemployment is data-oriented and empirical, his original and influential work on economic growth is mainly theoretical. He has pursued the implications of increasing returns in economic development ("Industrialization and the Big Push," with A. Shleifer and R. Vishny, 1989, in the *Quarterly Journal of Economics*). He has successfully confronted a central question in economic growth since Malthus by showing how an economy can make the transition from high fertility, low growth, and low human-capital intensity to low fertility, high growth, and high levels of educational attainment ("Human Capital, Fertility, and Economic Growth," with G. Becker and R. Tamura, 1990, in the *Journal of Political Economy*).

Minutes of the Annual Meeting Chicago, IL January 4, 1998

The one hundred and tenth annual meeting of the American Economic Association was called to order by President Arnold Harberger at 6:20 P.M., January 4, 1998, in the Grand Ballroom of the Hyatt Regency Hotel. Harberger began the meeting by announcing that copies of the agenda and the printed reports of various officers of the Association were available.

The first item on the agenda was consideration of the minutes of the previous annual meeting as published in the *Papers and Proceedings* issue of the *American Economic Review* (May 1997, p. 467). No corrections were offered, and the minutes were accepted without dissent as published.

The next items on the agenda were the reports of the Secretary (John J. Siegfried), Treasurer (C. Elton Hinshaw), Editor of the *American Economic Review* (Orley Ashenfelter), Editor of the *Journal of Economic Literature* (John Pencavel), Editor of the *Journal of Economic Perspectives* (Alan Krueger), and Director of *Job Openings for Economists* (Hinshaw). Each discussed his written report, which was available to the members prior to the meeting and is published elsewhere in this journal.

The Secretary announced that two changes in the Association's bylaws were APPROVED by a vote of the membership during 1997. The first change amends Article IV, paragraph 2 to permit the Executive Committee's spring meeting to be held no later than April 30 (rather than March 30); the second change amends Article IV, paragraph 3 providing that the slate of nominees selected by the Electoral College be announced to the membership no later than June 1 (rather than by the date of publication of the June issue of the *American Economic Review*).

The Treasurer reported a proposed budget that anticipates an overall deficit for 1998 of

\$299,000. Net worth at the beginning of 1998 is expected to exceed budgeted expenditures by over \$3 million, so the projected deficit can be accommodated.

The Editor of the *American Economic Review* reported that production delays had caused the size of several 1997 issues to be smaller than usual, but that more pages would be published in 1998 to compensate.

Before the Editor of the *Journal of Economic Literature* reported, it was VOTED unanimously that the Association express its appreciation to Pencavel for his dozen years of outstanding service to the Association as *JEL* Editor. Pencavel thanked the Association for the opportunity to edit the *Journal*. He announced that the new Editor is John McMillan, and reported that the transition of the *JEL*'s office from Stanford to the University of California at San Diego was proceeding smoothly.

The Editor of the *Journal of Economic Perspectives* announced that the *JEP* had just completed its tenth full year of publication. The recent change to encourage and publish more correspondence will be continued. Krueger also announced that, although not included in his written report, John Taylor and Timothy Bresnahan have been appointed as new members of the Board of Editors of the *JEP* with terms expiring in 2000.

The Editor of *Job Openings for Economists* reported that although *JOE* is now available free of charge on the Internet, there are still about 1,000 paid subscribers to the print version.

There being no questions about the reports and no further business before the assembly, the meeting was adjourned.

Respectfully submitted,
JOHN J. SIEGFRIED, *Secretary*

Minutes of the Executive Committee Meetings

Minutes of the Meeting of the Executive Committee in San Francisco, CA, March 7, 1997.

The first meeting of the 1997 Executive Committee was called to order at 10:12 A.M., March 7, 1997, in the San Francisco Hilton and Towers in San Francisco, CA. Members present were Orley Ashenfelter, Rebecca Blank, Ronald Ehrenberg, Robert Fogel, Victor Fuchs, Arnold Harberger, C. Elton Hinshaw, Alan Krueger, Anne Krueger, Glenn Loury, Rachel McCulloch, John Pencavel, John Roemer, Paul Romer, and John Siegfried. Attending for parts of the meeting were members of the Nominating Committee: Sebastian Edwards, Ronald Jones, Walter Oi, Nancy Rose, Amartya Sen, and Robert Topel. Also attending for parts of the meeting were members of the Honors and Awards Committee: Olivier Blanchard, Timothy Bresnahan, Avinash Dixit, Claudia Goldin, Dale Jorgenson, and Finis Welch. Eric Hanushek reported on behalf of the Search Committee for a New Editor of the *Journal of Economic Literature*.

Harberger opened the meeting by asking for approval of the minutes of the previous meeting (January 3, 1997) which had been circulated previously. The minutes were corrected to include the fact that Alan Krueger had attended and were otherwise approved as written.

Report of the Ad Hoc Committee on President-elect Nominations (Harberger for Solow). At the request of several members of the Executive Committee, in January 1997 Harberger appointed an ad hoc committee to examine the process whereby the Nominating Committee reports its recommendations for President-elect to the Executive Committee. The ad hoc committee consisted of Robert Solow (Chair), Zvi Griliches, and William Baumol. The committee recommended that the Nominating Committee submit a single name to the Electoral College (which consists of the Nominating Committee plus the voting members of the Executive Committee) and freely mention the names of a (very) few of the leading alternatives, together with the reasons for the Nominating Committee's ultimate choice.

Discussion of the ad hoc committee's recommendation revealed a broader concern about the procedure for nominating the President-elect than simply the Nominating Committee's procedure for reporting to the Electoral College. The Executive Committee decided to accept the report of the ad hoc committee and advise the Nominating Committee to follow the recommended procedure. It was then VOTED to appoint a new ad hoc committee to study more broadly whether there are reasonable ways of improving current nominating procedures for Association officers within the existing bylaws. It was recommended that the committee consult widely in its deliberations but necessarily talk with existing Executive Committee members about the issue. The new committee is expected to report at the January 1998 meeting of the Executive Committee.

Report of the Honors and Awards Committee (Jorgenson). Acting together as an electoral college, the Honors and Awards Committee and the Executive Committee VOTED to elect Kevin M. Murphy as recipient of the John Bates Clark Medal and Partha Dasgupta, Roger Guesnerie, Andreu Mas-Colell, and Stephen Nickell as Foreign Honorary Members.

The Honors and Awards Committee reported that it had solicited nominations for the Clark medal from all members of the Executive Committee, from past Clark medalists, and from the chairs of over a hundred university departments of economics.

Report of the Search Committee for a New Editor of the JEL (Hanushek). Hanushek reported that the committee, consisting of himself (as Chair), Robert Baldwin, Donald Brown, Marjorie McElroy, John Quigley, Jennifer Reinganum, Richard Rogerson, Sherwin Rosen, and John Siegfried (ex officio) recommended the appointment of John McMillan as the Editor of the *Journal of Economic Literature*, to succeed John Pencavel in January 1998. He reported that the committee advertised the position in *JOE* and sought additional candidates directly. Interested candidates were asked to describe their vision for the *JEL*. From these, John McMillan was se-

lected. The committee expressed confidence that McMillan's breadth of interest, ranging from game theory to emerging market economies, and his ideas for building on the *JEL*'s solid foundation would insure the *JEL*'s continued success.

On behalf of the Association, Harberger thanked John Pencavel for his fine service as *JEL* Editor. Harberger also thanked the Search Committee for completing its work quickly and identifying a highly qualified new editor. The Executive Committee then VOTED to publish the report of the Search Committee in the *Papers and Proceedings*.

Committee on Journals (Blank). Blank reported on the Committee's activities for Thomas Schelling, the Chair. The Committee met with the current editors of the Association's three journals in Chicago in February 1997. It expects to have a written report for consideration at the January 1998 Executive Committee meeting.

The 1998 Program (Fogel). Fogel announced that he was well into the process of organizing the program. He had received over 520 papers and suggestions for sessions that he planned to organize into 80 sessions of contributed papers distributed among fields in proportion to the distribution of submissions. He had organized 31 invited sessions so far. Although he elected to have no overall theme for the program, it emphasizes Asia, intergenerational equity issues, and poverty.

Report of the Nominating Committee (Sen). Sen, who chaired the Committee, reported the following nominations for the indicated offices: Vice-President—Robert Barro, Edward Lazear, June O'Neill, and Robert Pollak; Executive Committee—Timothy Bresnahan, Angus Deaton, Nancy Folbre, and Laurence Kotlikoff. The Nominating Committee and the Executive Committee, acting together as an electoral college, VOTED to nominate D. Gale Johnson as President-elect, and Martin Bronfenbrenner and Gordon Tullock as Distinguished Fellows.

Report of the Editor of the American Economic Review (Ashenfelter). Acting on the Editor's recommendation the Executive Committee VOTED to reappoint Dennis Epple as Co-Editor with a term ending May 31, 2000, and to re-appoint David Baron, Gary Solon,

W. Kip Viscusi, and Carl Walsh and to appoint Gordon Hanson, Deirdre McCloskey, Peter Reiss, and David Weil as members of the Board of Editors of the *AER*, each with a term ending March 31, 2000. There followed a general discussion about the growing difficulty of securing referees for submitted manuscripts.

Report of the Editor of the Journal of Economic Literature (Pencavel). Pencavel reported that the *JEL* has been considering the merits of putting the *Journal* on the Internet via its own website. The difficulty is that no one has yet figured out how to price it appropriately. He also noted that 59 journals would soon be eliminated from the set whose contents are listed in the printed version of the *JEL*. The electronic version of the listing includes 529 journals; 203 are in the printed version. The criteria for inclusion in the printed list include a heavy weight on citations. As fewer journals are included among those in the printed version of the *JEL*, and as the criteria have become known, the demand for inclusion in the printed list has intensified.

Report of the Editor of the Journal of Economic Perspectives (Krueger). Krueger reported that the *JEP* has instituted term limits for the Associate Editors and for the Advisory Board.

Report of the Director of Job Openings for Economists (Hinshaw). Hinshaw reported that the February issue was published and that subscriptions to the printed version of *JOE* continue to decline, no doubt as a consequence of *JOE* being available on the Internet for no charge.

Report of the Secretary (Siegfried). After welcoming Fogel, Loury, McCulloch, Romer, and Richard Freeman (who could not attend) as new members of the Executive Committee and promising a seamless transition from his predecessor, Hinshaw, Siegfried reviewed the schedule for sites and times of future meetings: Chicago, January 3–5, 1998 (Saturday, Sunday, and Monday); New York, January 3–5, 1999 (Sunday, Monday, and Tuesday); Boston, January 7–9, 2000 (Friday, Saturday, and Sunday); and New Orleans, January 5–7, 2001 (Friday, Saturday, and Sunday). As usual, the Executive Committee will meet one day prior to the beginning of the regular meeting. He reported that the process of

selecting a site for 2002 would begin soon. Locations under consideration include Atlanta, San Antonio, San Diego, San Francisco, and Washington, DC.

Seven thousand and fifty-eight people registered at the 1997 meeting in New Orleans. The previous meeting in New Orleans (1992) attracted 6,831. Registration for the 1996 meeting in San Francisco was 7,320. Fifty-two other associations, societies, and organizations met with us; 538 scholarly sessions were organized, and 179 "events" (lunches, cocktail parties, committee meetings, breakfasts, workshops, ship removals, etc.) were scheduled. Siegfried announced that a survey of members' preferences about meeting-site locations would be conducted in 1997.

A "Survey of Members" is scheduled for publication in December 1997. Members receive a print copy of the "Survey" at no additional charge. The biographical portion of the "Survey" is also installed on the Internet. The ensuing discussion of the merits of follow-up mailings of the questionnaire for the "Survey" took note of the Association's experience with an excellent response to first mailings for past surveys and the high cost of follow-up mailings.

At its March 1996 meeting, the Executive Committee asked the Secretary to do a study of Association membership. The Committee's concern centered more on membership than on Association revenues. Siegfried reported that he had compared AEA membership to a list of economists in the departments of economics at about 850 American colleges and universities. The study thus pertains only to U.S. academic economists; it excludes economics graduate students, foreign economists, economists employed outside academe, academic economists employed outside departments of economics, and emeritus economics faculty.

The results indicate that a little more than 54 percent (probably close to 60 percent) of academic economists are members of the Association. Membership percentages by rank are: full professors, 58 percent; associate professors, 49 percent; and assistant professors, 56 percent. Seventy percent of research university faculty, 59 percent of doctoral university faculty, 57 percent of selective liberal-arts college faculty, and 40 percent of comprehen-

sive university faculty are members. Membership varies across institutions, ranging from 54 percent to 92 percent among prominent research universities. Membership rates among assistant professors aggregated by the university at which they earned their terminal degrees range from 44 percent to 87 percent for programs with 25 or more graduates in the sample. There appears to be no trend in membership by age (using rank as a proxy).

The Secretary reported that the Association is participating in a joint Sloan Foundation/NSF-funded study of the labor market for Ph.D.'s in 12 natural- and social-science disciplines. The project is coordinated by the Commission on Professionals in Science and Technology. The pilot survey in economics will be conducted in 1997-1998. It will sample economists who earned their Ph.D.'s between July 1, 1996, and June 30, 1997. It will ask about their employment and earnings in October 1997, and their experiences in the job market.

On behalf of Susan Collins, chair of the Committee on the Status of Minority Groups in the Economics Profession, Siegfried reported that the National Science Foundation has awarded \$100,000 per year for three years (1997-1999) for support of the Association's Summer Program for Minority students, currently located at the University of Texas. Collins and Don Fullerton, Director of the program at Texas, indicated that the 1997 program could be managed with the NSF support plus the \$50,000 of AEA support approved for the Summer Program by the Executive Committee at its January 1997 meeting. Collins relayed that the Executive Committee would have to revisit the issue of the Minority Student Summer Program at its January 1998 meeting because of the Program's future funding uncertainty and the approaching end of the commitment from the University of Texas in 1998.

The Secretary concluded his report by noting that the Committee needed to establish the date and place of the 1998 Spring Meeting. Because of deadlines established in the Association's bylaws concerning the election process, March 31 is currently the latest date possible for the meeting. The constraint on the meeting date arises from the necessity to notify the membership of the slate of candidates for

Association offices in sufficient time to allow additional candidates to be nominated by petition. The current bylaws require the "Secretary [to] inform all members of the Association of the actions of the Nominating Committee and the Electoral College not later than the June issue of the *American Economic Review*." Siegfried noted that for nine years the membership has been notified of the slate of nominees in the Notes section of the Spring *JEP*, as that feature was transferred from the *AER* in 1988. He suggested that notification could be accomplished more quickly by using the Association's Home Page on the Internet, supplemented by mail or fax replies to requests directed to the Association office. He noted that those Association members (currently 746) who elect not to receive the *JEP* do not receive notification under current procedures. Notifying members of the slate of nominees by means of the Home Page would allow the Spring Meeting of the Executive Committee to be held in April, which would space the Executive Committee's two annual meetings better, would alleviate a peak load problem in auditing the Association's financial records, and would decrease the likelihood of bad weather impeding travel to the Spring Executive Committee meeting. The Executive Committee VOTED to submit to the membership a proposed change in the bylaws that alters from March 31 to April 30 the specified date by which the Nominating Committee must report to the Executive Committee, and specifies notification of the slate of nominees to the membership be completed by June 1 rather than by the publication date of the June issue of the *AER*. In addition to posting on the Home Page, a note in the *JEP* will give instructions for requesting a slate of nominees from the Association's office. The Executive Committee selected San Francisco as the site for the Spring 1998 Executive Committee meeting.

Report of the Treasurer (Hinshaw). The Treasurer distributed the Association's audited financial statements for 1996. The "Statements of Revenues and Expenses" showed an operating deficit of \$331 thousand for 1996, up from \$314 thousand the previous year. After taking into account recognized investment income of \$390 thousand, a surplus of \$59 thousand resulted. The operating loss was

\$374 thousand less than budgeted; revenues were \$179 thousand higher, and expenses were \$159 thousand lower than budgeted amounts. Revenues from SilverPlatter exceeded budget expectations by \$251 thousand. Investment income realized exceeded the amount budgeted by \$46 thousand. Consequently instead of an overall budgeted deficit of \$361 thousand, the Association realized a surplus of \$59 thousand. The ratio of net worth at December 31, 1996, to 1997 budget expenses is 1.63.

Other Business (Harberger). Harberger reported his interest in promoting further consideration within the discipline of the Report of the Committee on Graduate Education in Economics, chaired by Anne Krueger. Krueger added her view that it would be useful to update some of the labor-market information that was collected by that Committee.

There being no further business to conduct, it was VOTED to adjourn at 4:50 P.M.

Minutes of the Meeting of the Executive Committee in Chicago, IL, January 2, 1998.

The second meeting of the 1997 Executive Committee was called to order at 10:16 A.M. January 2, 1998, in the Hyatt Regency Hotel in Chicago, IL. Members present were Orley Ashenfelter, Ronald Ehrenberg, Robert Fogel, Richard Freeman, Arnold Harberger, C. Elton Hinshaw, Alan Krueger, Anne Krueger, Glenn Loury, Rachel McCulloch, John Pencavel, John Roemer, Paul Romer, John Siegfried, and Barbara Wolfe. Attending as guests were Angus Deaton, D. Gale Johnson, and June O'Neill, recently elected officers of the Association, and John McMillan, recently appointed editor of the *Journal of Economic Literature*. Attending for parts of the meeting for the purpose of giving committee reports were: Robin Bartlett, reporting on behalf of the Committee on the Status of Women in the Economics Profession; Susan Collins, reporting on behalf of the Committee on the Status of Minority Groups in the Economics Profession; Walter Oi, reporting on behalf of the Ad Hoc Committee on Nominating Procedures; and Thomas Schelling, reporting on behalf of the Ad Hoc Committee on Journals.

Harberger opened the meeting by asking for approval of the minutes of the previous

meeting (March 7, 1997) which had been circulated in advance. The minutes were approved as written.

Report of the Secretary (Siegfried). Before giving his report, Siegfried urged Executive Committee members to attend the annual business meeting of the Association. He announced that Rebecca Blank had resigned from the Executive Committee in October 1997 in order to accept an appointment to the President's Council of Economic Advisers. He further noted that this was the last meeting for Victor Fuchs, Richard Freeman, Glenn Loury, and John Roemer as members of the Executive Committee. On behalf of the Association, he thanked them for their service. The Secretary also noted that the staff of the Association in Nashville had been very helpful during his first year as Secretary of the Association.

Siegfried then reviewed the schedule for sites and dates of future meetings: New York, January 3–5, 1999 (Sunday, Monday, and Tuesday); Boston, January 7–9, 2000 (Friday, Saturday, and Sunday); and New Orleans, January 5–7, 2001 (Friday, Saturday, and Sunday). As usual, the Executive Committee will meet one day prior to the beginning of the regular meeting.

Atlanta, San Antonio, San Diego, San Francisco, and Washington, DC, had been contacted as possible sites for the 2002 annual meeting. San Diego is not available, San Antonio currently does not have sufficient hotel rooms in a suitable configuration to accommodate the meetings, and San Francisco submitted a bid that was not competitive. It was VOTED to authorize the Secretary to negotiate a contract with either Atlanta or Washington, DC for the 2002 meeting. Suggestions for the site of the 2003 meeting were solicited.

Along with the 1997 election of officers, the membership had been polled as to their preferences for meeting sites. Respondents were asked to identify their five most preferred sites; 4,634 preferences were returned. The results of the poll are:

San Francisco	2,494
Washington, DC	2,189
New York	1,854
Boston	1,847
Seattle	1,650

San Diego	1,582
New Orleans	1,509
Chicago	1,283
Denver	965
San Antonio	952
Orlando	938
Atlanta	927
Philadelphia	909
Las Vegas	894
Los Angeles	772
Anaheim	506.

The meetings of 1996–2001 are in sites ranked 1, 3, 4, 7, and 8. Sites ranked 2 and 6 are under consideration for 2002 and 2003. The fifth ranked site, Seattle, cannot presently accommodate the meeting in a single location with hotels within walking distance but could be viable if one additional judiciously located convention hotel were built.

Siegfried reported that members' preferences figure significantly into the selection of a meeting site but are not the sole criterion. Hotel room rates, the hotel configuration and meeting room capacity, airline accessibility, and the probability of weather disrupting the meeting also influence site selection. Among the members' top dozen preferences, Atlanta, Chicago, and New Orleans rate highest on the criteria other than members' preferences.

There followed a discussion of the optimal size of the annual meeting. Over 6,100 rooms were booked in Chicago, a new record by a substantial margin. The combined growth of the meeting, desire to use hotel (rather than convention-center) meeting facilities, and the current three-day length of the meeting constrain site options. A preference was indicated to attempt to control the rapid growth in sessions, but simultaneously to encourage attendance. There was no enthusiasm for adding a fourth day of sessions or for moving the meeting to a convention center.

The number of sessions has grown 27 percent, from an average of 427 during 1986–1990 to 543 during 1996–1998. Average registration has increased by only 3 percent, from about 7,100 to 7,300 over the same period. Siegfried noted that 1999 is a propitious time to reduce the number of sessions, because recent renovations of the co-headquarters hotel in New York have reduced the number of meeting

rooms so much that the 549 sessions scheduled for Chicago could not be accommodated in New York. He agreed to coordinate a reduction in sessions among the 51 other associations that meet simultaneously with the Association under the ASSA designation.

The Secretary reported that the transition of the *Journal of Economic Literature* editorship from John Pencavel at Stanford to John McMillan at the University of California-San Diego is in progress. Both editorial offices will operate simultaneously during the first quarter of 1998, after which the Stanford office will close.

A new "Survey of Members" has been published and was mailed to members in December 1997. It contains brief bibliographic facts about members as well as postal and e-mail addresses and fax and telephone numbers. It will be posted on the Internet by February 1998. The next "Survey of Members" is scheduled for 2001. Following the traditional pattern, a printed "Telephone Directory" would normally be published in 1999. Because the regularly updated Association membership list is available on the Internet (including telephone, fax, and e-mail contacts), a printed "Telephone Directory" in 1999 may not be cost-effective. Members will be polled regarding their need for a printed "Telephone Directory" along with the 1998 ballot for election of officers.

The Amendment to Article IV, paragraph 2 of the Association's Bylaws that permits the spring Executive Committee meeting to be held as late as April 30 (vis-à-vis March 30) was passed by a vote of the membership. The first meeting of the 1998 Executive Committee is scheduled for Friday, April 17, 1998, in San Francisco.

An Executive Committee meeting date in April does not allow sufficient time to announce the slate of nominees for Association offices in the Spring issue of the *Journal of Economic Perspectives*. The Amendment to Article IV, paragraph 3 of the Bylaws providing that the slate of nominees be announced to the Association membership by June 1 was also passed by a vote of the membership. (It replaces a provision requiring the membership to be notified by the date of publication of the June issue of the *American Economic Review*.)

The Secretary will announce the slate of nominees on the Association's Internet Home Page and will provide it in writing upon request.

Since 1989, members have had the option of not receiving one of the three journals. As of November 1997, 2,376 members have elected this option: 770 chose not to receive the *AER*; 865, the *JEL*; and 732, the *JEP*. Nine members have paid their dues and elected to receive none of the journals.

Last year John Coffee, the Association's General Counsel since 1992, indicated a desire to pass the opportunity to represent the Association to a successor. It was VOTED to appoint Terry Calvani, a partner in the San Francisco law firm of Pillsbury, Madison & Sutro to a three-year term as the Association's General Counsel. Calvani is a former Professor of Law at Vanderbilt University and a former Federal Trade Commissioner. Although he specializes in antitrust law, he is located in Pillsbury, Madison & Sutro's Washington DC office, which specializes in intellectual property law.

It was VOTED to express the Association's gratitude to John Coffee for his service as General Counsel since 1992.

Siegfried reported that the Association is participating in an NSF-supported survey of the labor market for economics Ph.D.'s. Economics is one of 12 natural- and social-science disciplines conducting comparable surveys. The purpose is to inform current and prospective Ph.D. students about the employment experiences of recent doctoral graduates. To encourage participation in the survey, a \$40 certificate toward a new or renewal regular membership in the Association is provided to respondents. The certificates may also serve to introduce to the Association economists who might otherwise not join. The distribution and redemption of the certificates and persistence of those who redeem them will be monitored.

A discussion of the merits of requiring Association membership of authors of abstracts submitted for the Contributed Papers portion of the annual program then ensued. It was VOTED to require that at least one author of each individual paper proposed for the Contributed Papers portion of the annual meeting program be an Association member. This requirement applies to all papers on sessions

proposed for joint AEA sponsorship by other Allied Social Science associations. It is to take effect for the January 2000 meetings in Boston.

Report of the Editor of the American Economic Review (Ashenfelter). Ashenfelter commented on his written report which is published elsewhere in this issue of the *AER*. He noted that the *AER* published fewer pages in 1997 than in the previous few years because of a transition in production editors. The backlog of accepted articles has increased, but it will be reduced by publishing relatively more pages in 1998. He also noted that it is getting increasingly difficult to secure referees, which is extending the delay in reaching a decision on submitted articles. To increase member awareness of the review process, the *AER* will begin to publish portions of the Editor's annual report in its regular issues as well as in the *Papers and Proceedings*. There followed a discussion of how to accelerate the refereeing process. No conclusion was reached.

Acting on Ashenfelter's recommendation, the Committee VOTED to approve the reappointment of Adam Jaffe, Theodore Bergstrom, Charles Brown, Don Fullerton, Robert Moffitt, Jennifer Reinganum, Andrew Schotter, and Curtis Taylor and the appointment of Susanto Basu, Francine Blau, Martin Gaynor, Randall Wright, Tracy Lewis, and Robert Staiger to terms on the *AER* Board of Editors. Jaffe, Basu, Blau, Gaynor, and Wright's terms expire in 2000; the terms of the others expire in 2001.

Report of the Editor of the Journal of Economic Literature (Pencavel). Pencavel reviewed the main points of his written report, which is published elsewhere in this issue of the *AER*. It was VOTED to approve Pencavel's recommendation to reappoint Alan Auerbach, Peter Howitt, Peter Reiss, F. M. Scherer, and John Whitaker, and to appoint Lewis Evans, James Levinsohn, Glenn Loury, Jennifer Reinganum, Michael Rothschild, Suzanne Scotchmer, Hans-Werner Sinn, Hal Varian, and John Pencavel to the *JEL* Board of Editors for three-year terms.

It was also VOTED to express on behalf of the entire membership the Committee's deep appreciation to Pencavel for his 12 years of service to the economics profession as Editor

of the *JEL*. Pencavel thanked the Executive Committee and the Association for the opportunity to edit the *JEL*.

Report of the Editor of the Journal of Economic Perspectives (Alan Krueger). After reviewing his written report, which is published elsewhere in this issue of the *AER*, Krueger reported that with the Spring 1997 issue, the *JEP* had completed ten years of publication. Although the *JEP* consists mostly of invited papers, it receives about 200 unsolicited manuscripts annually. Each is considered carefully. A few eventually appear in the *Journal* in revised form, but most are inappropriate for the *JEP*. The recent policy to include more correspondence will continue. It was VOTED to approve Krueger's recommendation to reappoint Francine Blau and Anne Case and to appoint Timothy Bresnahan and John Taylor to the Board of Editors of the *JEP* for three year terms.

Report of the Director of Job Openings for Economists (Hinshaw). Hinshaw reviewed his written report, published elsewhere in this issue of the *AER*. He noted that the decline in paid subscriptions to the print version of *JOE* had slowed and seems to be stabilizing near 1,000. *JOE* is available on the Internet free of charge several days after the print copies are mailed.

The 1998 Program (Fogel). Fogel thanked members of his program committee and the staff in Nashville and expressed his special appreciation for the efforts of Karen Brobst in Chicago, who did a large part of the administrative work organizing the 1998 program.

The 1999 Program (Johnson). Johnson indicated that he did not plan to have a single theme for the 1999 program but, rather, would gather together collections of sessions on income-distribution data deficiencies, transitions in the economies of Central and Eastern Europe, China's income distribution, and a series of policy issues such as fertility controls, trade embargoes, and the travails of East Asian economies. He announced that Stanley Fischer had agreed to be the speaker at the joint AEA/AFA luncheon.

Report of the Committee on the Status of Minority Groups in the Economics Profession (Collins). After reviewing her report, published elsewhere in this issue of the *AER*,

Collins reported that the CSMGEP's \$700,000 proposal to the MacArthur Foundation for support over four years had been approved. The MacArthur funds provide partial support for continuation of the CSMGEP's Summer Program at the University of Texas and also support a new pipeline initiative. The pipeline project consists of an outreach program to minority undergraduates and a mentoring program for minorities in Ph.D. programs. The MacArthur support has allowed CSMGEP to "stretch out" its grant from NSF in order to extend the Summer Program through 2001. It was VOTED to congratulate Susan Collins and CSMGEP for securing support from the MacArthur Foundation, to extend the term of the Summer Program at the University of Texas for two additional years (through summer 2000), to authorize CSMGEP to search for a director of the outreach and mentoring components of the pipeline project, to approve a \$50,000 allocation from the Association toward the 1998 Summer Program, and to approve a \$25,000 allocation from the Association during 1998 toward advance expenses of the 1999 Summer Program.

There followed an extensive discussion of means to evaluate the effectiveness of the Summer Program, including the appropriate criteria for measuring the "success" of the program, the challenge of identifying an appropriate control group against which to measure success, and consideration of the efficacy of alternative approaches to encourage the development of minority economists. The need to "match" minority Ph.D. candidates with appropriate graduate schools and training was also discussed.

Report of the Committee on the Status of Women in the Economics Profession (Bartlett). After reviewing her written report, published elsewhere in this issue of the *AER*, Bartlett announced that CSWEP was celebrating its 25th anniversary at this annual AEA meeting. She reported that the percentage of women at assistant professor ranks had grown from 7 percent in 1974, soon after CSWEP was formed, to 24 percent in 1996, and the percentage of full professors who are women had grown from 2 percent in 1974 to 8 percent in 1996. Bartlett noted that the commemorative Fall 1997 CSWEP Newsletter includes a history

of CSWEP, a new mission statement for CSWEP, and articles discussing CSWEP's heritage, the present situation for women economists in academe and business, and CSWEP's plans for the future.

CSWEP is conducting a two-day NSF-supported mentoring workshop for 40 young women economists in Chicago immediately after the regular meeting concludes. The purpose of the workshop is to accelerate the progress of women economists through the academic ranks. One objective of the workshop is to identify impediments to women economists' progression from assistant to full professor. The workshop will be repeated at regional economics association meetings throughout the year.

Bartlett reported that CSWEP has received two opportunities to establish awards. The first would be supported by a \$20,000 donation from William Zame in memory of his late wife, Elaine Bennett, who was a prominent economist. Zame would like to establish a prize that would bring attention to the professional accomplishments of a young woman economist. Second, CSWEP would like to establish an award that would recognize the efforts of persons who have furthered the careers of women economists in academe, business, and government in honor of Carolyn Shaw Bell, who has herself mentored over 50 women economists.

The advantages and disadvantages of establishing awards for a specific group of economists was discussed at length. After it was clarified that the awards would not be construed as AEA prizes, and that no monetary emolument would be awarded, it was VOTED to authorize CSWEP to establish both the Bennett and Bell awards as CSWEP awards.

Report of the Ad Hoc Committee on Journals (Schelling). Schelling reported that his Committee had a final meeting in December and had prepared a draft report. He promised the Executive Committee a final report at the April 1998 meeting. Schelling reported that the Committee's deliberations had been informed by a survey of members about the Association's journals. The Executive Committee requested that Schelling try to provide them with some cross-tabulations of survey results by the extent of respondents' own research activity.

Report of the Ad Hoc Committee on Nominating Procedures (Oi). Oi reviewed the background for the Ad Hoc Committee report. At the January 1997 meeting, the Executive Committee asked then President-elect Harberger to establish a committee to study whether only one name should be brought forward to the Electoral College by the Nominating Committee for the office of President-elect. A committee chaired by Robert Solow (William Baumol and Zvi Griliches, members) was appointed and reported in February 1997. During the discussion of the report at its March 1997 meeting, the Executive Committee decided to conduct a more extensive review of Association nomination procedures. President Harberger responded in April 1997 by appointing an Ad Hoc Committee on Nominating Procedures consisting of Robert Baldwin, Claudia Goldin, Allen Kelley, Walter Oi (Chair), John Pencavel, and James Poterba.

The Ad Hoc Committee was charged with examining the composition of the Nominating Committee, how much of the Nominating Committee's deliberations should be shared with the Executive Committee when they meet jointly as the Electoral College, the voting procedure used by the Electoral College, the relationship of the nominating process for Distinguished Fellows to that for President-elect, and other issues relevant to these matters.

Article IV, Section 2 of the Association Bylaws stipulates that before October 1 of each year the President-elect shall appoint a Nominating Committee for the following year. This Committee shall be chaired by a former officer of the Association. The President-elect must appoint at least five other members of the Association to serve on the Nominating Committee. Traditionally the third past-President of the Association (who has just completed his/her term on the Executive Committee) chairs the Nominating Committee. From 1989 to 1993, the Nominating Committee contained six members plus the Chair; since 1994 it has contained seven members plus the chair. For at least a decade, the President-elect has followed a tradition of appointing one "holdover" member from the previous Nominating Committee to provide institutional memory.

At least two members of the Ad Hoc Committee interviewed each member of the current Executive Committee about the nomination process. The Committee also collected information about procedures used to nominate officers by five other academic associations.

The Ad Hoc Committee concluded that the size of the Nominating Committee should be at least five and no more than seven (plus the Chair). Because the statutory minimum is five, and the current tradition is to select seven, no change was recommended in the size of the Nominating Committee. Nor was a change in the tradition of appointing one "hold-over" member to the Nominating Committee recommended.

The Ad Hoc Committee recommended and it was VOTED that the President-elect should appoint a Nominating Committee consisting of: (1) the third past-President of the Association as chair; (2) two "set-aside" appointments, one of a former Vice-President and one of a former member of the Executive Committee, each of whom has just completed his/her term of office, preferably chosen by lot; and (3) at least three other Association members. The Ad Hoc Committee also recommended and it was VOTED to instruct the Nominating Committee to bring to the Electoral College at least two names for the office of President-elect, rank-ordered if they wish. No change was recommended for the Nominating Committee's procedure to select the slate of candidates for the offices of Vice-president and Executive Committee Member. The Ad Hoc Committee recommended and it was VOTED to instruct the Nominating Committee to bring to the Electoral College at least three nominations (rank-ordered if they wish) from which, after discussion, a maximum of two shall be selected for the honor of Distinguished Fellow of the Association.

The Ad Hoc Committee also recommended and it was VOTED that the Nominating Committee should present a rationale for their choices and ranking of candidates for President-elect and Distinguished Fellow to the Electoral College orally.

A discussion of the implications of these changes then ensued. It was concluded that the Vice-presidents and Executive Committee mem-

bers leaving office after the annual business meeting in early January of year x were those eligible for the "set-aside" appointments for the Nominating Committee for year x (usually meeting in April of said year), and that all of these changes in nominating procedures should be implemented as soon as possible. It was decided that voting among candidates for Distinguished Fellow should follow a two-step procedure. The first step should be a vote on the acceptability of each candidate. That should be followed by a vote wherein each member of the Electoral College rank-orders the candidates found acceptable in the first step, the "rank votes" then being summed (lower numbers being a higher preference), and the candidate(s) with the lower total selected until the positions are filled or nominees are exhausted, whichever comes first.

There followed a brief discussion of the means by which the Nominating Committee identifies candidates for offices. In addition to suggestions from themselves, the Nominating Committee usually communicates with the chair of the previous Nominating Committee. Each year the Secretary also provides the Nominating Committee with a list of Association members who are named frequently on the annual ballot request for suggested nominees for Association offices.

Report of the Treasurer (Hinshaw). Hinshaw presented the proposed 1998 budget,

which is published elsewhere in this issue of the AER. He projected an operating loss of \$820 thousand, investment income of \$496 thousand, and an overall deficit for the year of \$324 thousand. Net worth at the beginning of 1998 was expected to exceed budgeted annual expenditures for 1998 by over \$3 million. The projected deficit could be accommodated.

Hinshaw noted that he had submitted a proposed budget showing an operating deficit every year he had been Treasurer, although in some of those years the Association had failed to achieve its goal of a deficit. The recent rise in the stock market coupled with the Association's policy of realizing a constant 5 percent of net assets as income means that for now the Association can handle deficit operating budgets. It was VOTED to approve the budget.

For the Good of the Association. The Secretary circulated tables reporting the number of undergraduate degrees awarded in each of the last six years at 111 U.S. colleges and universities. He noted that the precipitous decline that occurred between 1992-1993 and 1994-1995 had slowed in 1995-1996, and the trend finally turned upward (modestly) in 1996-1997.

There being no further business to conduct, it was VOTED to adjourn.

Respectfully submitted,
JOHN J. SIEGFRIED, *Secretary*

Report of the Secretary for 1997

Membership Survey. The Association published a new *Survey of Members* in December 1997. It contained mailing addresses, telephone numbers; fields of research interest, and a brief biographical sketch of each active member. That *Survey* is now available on the Internet. To access it, connect to gopher.eco.utexas.edu using Gopher or a World Wide Web browser such as Lynx or Mosaic. Select "Search the membership directory" from the ensuing menu.

Annual Meetings. The next annual meeting will be held in New York, NY, January 3–5, 1999. Placement Service will open for business one day earlier than the meetings. The schedule for subsequent meetings is Boston (2000) and New Orleans (2001).

Elections. In accordance with the bylaws on election procedures, I hereby certify the results of the recent balloting and report the actions of the Nominating Committee and the Electoral College.

The Nominating Committee, consisting of Amartya Sen (Chair), Michael J. Boskin, Sebastian Edwards, Ronald W. Jones, Walter Y. Oi, Nancy L. Rose, T. Paul Schultz, and Robert Topel submitted the nominations for Vice-Presidents and members of the Executive Committee. The Electoral College, consisting of the Nominating Committee and Executive Committee meeting together, selected the nominee for President-elect. No petitions were received nominating additional candidates.

President-elect

D. Gale Johnson

Vice-President

Robert J. Barro
Edward Lazear
June E. O'Neill
Robert A. Pollak

Executive Committee

Timothy F. Bresnahan
Angus S. Deaton
Nancy Folbre
Laurence J. Kotlikoff

The Secretary prepared biographical sketches of the candidates and distributed ballots last summer. On the basis of the canvass of ballots, I certify that the following persons have been duly elected to the respective offices:

President-elect (for a term of one year)

D. Gale Johnson

Vice-Presidents (for a term of one year)

Robert J. Barro
June E. O'Neill

Executive Committee (for a term of three years)

Angus S. Deaton
Laurence J. Kotlikoff

In addition, I have the following information:

Number of legal ballots	4,648
Number of invalid envelopes	341
Number of envelopes received after October 1	57
Number of envelopes returned	5,046

The proposed amendments to the bylaws were adopted; the bylaws as amended now read:

Article IV. DUTIES OF OFFICERS (in part)

Section 2 (Paragraph 2, Sentence 2).

The Nominating Committee for each year shall be instructed to present to the Executive Committee on or before April 30 a nominee for the President-elect and two or more nominations for each other elective office to be filled, except the presidency, all these nominees being members of the Association.

Section 2 (Paragraph 3, Sentence 1).

The Secretary shall announce the action of the Nominating Committee and the Electoral College to members of the Association no later than June 1.

Membership. The total number of members and subscribers is shown in Table 1.

Permission to Reprint and Translate. Official permission to quote from, reprint, or translate and reprint articles from the *American Economic Review*, *Journal of Economic Literature*, and the *Journal of Economic Perspectives* totaled 1,310 in 1997, compared to 1,626 in 1996. Upon receipt of a request for permission to reprint an article, the publisher or editor making the request is instructed to obtain the author's permission in writing and send a copy to the Secretary as a condition for official permission. The Association suggests that authors charge a fee of \$150, but they may charge some other amount, enter into a royalty arrangement, waive the fee, or refuse permission altogether.

TABLE 1—MEMBERS AND SUBSCRIBERS (END OF YEAR)

Class of membership	1993	1994	1995	1996	1997
Regular	18,276	17,979	17,878	17,649	18,339
Junior	2,642	2,612	2,640	2,554	2,424
Life	311	305	301	294	282
Honorary	37	34	38	41	40
Family	416	400	386	194	362
Complimentary	323	319	322	324	273
Total members	22,005	21,649	21,565	21,056	21,720
Subscribers	5,531	5,474	5,384	5,219	5,234
Total members and subscribers	27,536	27,123	26,949	26,275	26,954

Staff. My primary responsibility is to retain our loyal, gracious, and efficient staff. I am indebted to them for helping make my initiation into this job as easy as it has been. They are Norma Ayres, Marlene Hight, Kimberly Johnson, Cindy Jones, Dana Ragan, Kim Roberts, Violet Sikes, and Mary Winer.

Committees and Representatives. Listed below are those who served the Association during 1997 as members of committees or representatives. The year in parentheses indicates the final year of the term to which they were appointed. On behalf of the Association, I thank them for all their service.

AEA COMMITTEES 1997

Ad Hoc Committee on Journals

Thomas Schelling, Chair
Alan Auerbach
Rebecca Blank
William Darity, Jr.
Nathan Rosenberg

Ad Hoc Committee on Nominating Procedures

Walter Oi, Chair
Robert E. Baldwin
Claudia Goldin
Allen C. Kelley
John Pencavel
James M. Poterba

Budget Committee

Elton Hinshaw, Chair
Rebecca Blank (1997)
Arnold C. Harberger (1997)
Ronald G. Ehrenberg (1998)

Robert W. Fogel (1998)
Rachel McCulloch (1999)

Census Advisory Committee

Michael Gort (1997)
Ellen Dulberger (1997)
William C. Dunkelberg (1997)
Lee Lillard (1997)
Robert Willis (1997)
Ernst R. Berndt (1998)
Ariel Pakes (1998)
Roger R. Betancourt (1999)
Frederic M. Scherer (1999)

COSSA-Liaison Committee

Charles Plott, Chair (1998)
Henry Aaron (1997)
John Taylor (1997)
Robert Haveman (1999)

Committee on Economic Education

Michael K. Salemi, Chair (1999)
Charles M. Kahn (1999)
Thomas H. Tietenberg (1999)
Cecilia Conrad (1997)
William T. Gavin (1997)
Hal Varian (1997)
Paul Krugman (1998)
Michael Watts (1998)
John Taylor (1999)
William E. Becker, Jr., ex officio

Committee on Electronic Publishing

Daniel McFadden, Chair
Malcom Getz
Hal Varian
Orley Ashenfelter

John Pencavel
Alan Krueger

Finance Committee

Elton Hinshaw, Chair
Robert Eisner (1997)
Robert Dederick (1998)
Robert Hamad  (1999)

Committee on Honors and Awards

Dale W. Jorgenson, Chair (1997)
Beth Allen (1997)
Olivier Blanchard (1999)
Avinash Dixit (1999)
Finis Welch (1999)
Timothy Bresnahan (2001)
Claudia Goldin (2001)

1997 Nominating Committee

Amartya Sen, Chair
Michael J. Boskin
Sebastian Edwards
Ronald W. Jones
Walter Y. Oi
Nancy L. Rose
T. Paul Schultz
Robert Topel

*Committee on the Status of Minority Groups
in the Economics Profession*

Susan M. Collins, Chair (1998)
Lynn Burbridge (1997)
Barbara Robles (1997)
Warren Whatley (1997)
Alvin E. Headen, Jr. (1999)
Willene A. Johnson (1999)
Richard Santos (1999)

*Committee on the Status of Women in the Eco-
nomics Profession*

Robin Bartlett, Chair (1999)
Maureen Cropper (1997)
Daphne Kenyon (1997)
Kenneth Small (1997)
Hali Edison (1998)
Joyce P. Jacobsen (1998)
Olivia S. Mitchell (1998)
Susan Pozo (1998)
Catherine C. Eckel (1999)
Henry Stuart Farber (1999)
Arleen Leibowitz (1999)
Nancy Lin Rose (1999)
Arnold C. Harberger, ex officio
Joan G. Haworth, Membership Secretary

COUNCIL AND OTHER REPRESENTATIVES

*American Association for the Advancement of
Science, Section K, Social Economics and Po-
litical Sciences*

Joseph Newhouse (2001)

*American Association for the Advancement of
Slavic Studies*

John P. Hardt (1997)

American Council of Learned Societies

John J. Siegfried (1998)

*Consortium of Social Science Associations
(COSSA)*

John J. Siegfried

*Council of Professional Associations on Fed-
eral Statistics (COPAFS)*

Janet Norwood (1998)
Marc L. Nerlove (1999)

International Economic Association

Anne Krueger (1999)

National Bureau of Economic Research

John J. Siegfried (1999)

Social Science Research Council

Michelle J. White (1999)

REPRESENTATIVES OF THE ASSOCIATION ON VARIOUS OCCASIONS—1997

Inaugurations

Lee C. Bollinger, University of Michigan
Steffany G. Ellis
Mark G. Yudof, University of Minnesota
Jill Boylston Herndon

JOHN J. SIEGFRIED, *Secretary*

Report of the Treasurer for the Year Ending December 31, 1997

The proposed budget for 1998 in Table 1 projects an operating loss of \$820 thousand, investment income of \$496 thousand, and an overall deficit for the year of \$324 thousand. Net worth at the beginning of 1998 is expected to exceed budgeted expenditures by over three million dollars. The projected deficit can be accommodated.

Audited statements for 1997 will be published in the June issue of the *American Economic Review*.

I thank Norma Ayres, our accountant, and Mary Winer, the Administrative Director, for their valuable help and patience in assisting me in carrying out the duties of the Treasurer.

C. ELTON HINSHAW, *Treasurer*

TABLE 1—1998 BUDGET, AMERICAN ECONOMIC ASSOCIATION (THOUSANDS OF DOLLARS), PREPARED JANUARY 2, 1998

	First nine months (unaudited)		Full year		
	1996	1997	Actual 1996	Budgeted 1997	Budgeted 1998
REVENUES FROM DUES AND ACTIVITIES:					
Membership dues and nonmember subscriptions	\$1,496	\$1,550	\$2,012	\$2,128	\$2,208
<i>Job Openings for Economists</i> , Subscriptions	24	16	34	25	20
Advertising	96	109	133	165	155
Sales of <i>Index of Economic Articles</i>	63	14	76	115	96
Sales of publications, reprints	40	32	50	45	42
Sale of mailing list	38	53	47	62	65
Annual meeting	105	80	93	95	80
Royalties—Knight-Ridder	27	25	35	31	31
Royalties—SilverPlatter	416	448	798	714	773
Royalties—EconLit	38	43	51	43	43
Royalties—OCLC	32	45	75	86	91
Royalties—miscellaneous	21	21	26	25	28
AER submission fees	39	42	55	50	55
Sundry	7	4	24	35	25
Total Operating Revenue	2,442	2,482	3,509	3,619	3,712
PUBLICATION EXPENSES:					
<i>American Economic Review</i>	761	735	940	1,114	1,115
<i>Journal of Economic Literature</i>	1,040	1,067	1,423	1,460	1,547
<i>Journal of Economic Perspectives</i>	376	360	497	537	554
<i>Survey and Telephone Directory</i>	30	30	40	41	40
<i>Job Openings for Economists</i>	54	52	78	80	84
<i>Index of Economic Articles</i>	92	99	174	149	321
Subtotal	2,353	2,343	3,152	3,381	3,661
OPERATING AND ADMINISTRATIVE EXPENSES:					
General and administrative	407	426	503	605	623
Committees	80	98	124	150	168
Support of other organizations	59	63	78	80	80
Subtotal	546	587	705	835	871
Total Expenses	2,899	2,930	3,857	4,216	4,532
OPERATING GAIN (LOSS)	(457)	(448)	(348)	(597)	(820)
INVESTMENT GAIN (LOSS)	293	329	390	401	496
Surplus (Deficit)	(\$164)	(\$119)	\$42	(\$196)	(\$324)

Report of the Finance Committee

The Finance Committee of the American Economic Association met at the Chicago Club, Chicago, IL, at 11:45 A.M. on December 18, 1997. Present were, Robert Eisner, Robert Dederick (members of the committee), and C. Elton Hinshaw (Chairman of the Committee and Treasurer of the Association), John Siegfried (Secretary of the Association); and Robert McNeill, Scott W. Vogg, and Debbie Jansen (representing Stein Roe & Farnham, investment counsel for the Association).

In 1987, the committee reviewed recommendations presented by the AEA Committee on Indexing Association Funds concerning the long-term allocation of the Association's investment assets. As a result of that recommendation and the subsequent deliberation of the Finance Committee, it was agreed that the Association's portfolio comprise a combination of a S&P 500 Index Fund, Stein Roe & Farnham's specialty equity mutual funds, and a bond portion managed by Stein Roe & Farnham.

This restructuring took place at the end of June 1988. The current portfolio includes holdings in the Vanguard Index Trust Fund, as well as several special Growth Funds and an International Fund, under the supervision of Stein Roe & Farnham (SRF). The Fixed Income portion of the portfolio is currently invested in SRF's Money Market and Intermediate Term Government and Corporate taxable funds.

As a result of the aforementioned asset allocation restructuring, the overall performance of the Association's fund now reflects the combined efforts of Vanguard and Stein Roe & Farnham.

With respect to the calendar-year 1997 performance of the Association's portfolio, the total return of the account, including cash, bonds, and equity holdings, was approximately 22.7 percent.

Last year the Committee approved a 60–85-percent allocation to equities (including the Vanguard fund). In addition, it was decided that international equity exposure should range from 5 percent to 20 percent, and the minimum cash equivalent position should be 0–10 percent. This year the Committee reaffirmed the existing investment guidelines and agreed (not unanimously) that the current overweighted equity position should be maintained. The benchmark for performance is a portfolio consisting of 60 percent S&P 500, 12.5 percent EAFE, 22.5 percent Lehman Corporate/Government Intermediate Index, and 5 percent cash. This benchmark generated a total return of 20.5 percent for 1997.

Members can obtain a list of the assets in the portfolio by writing the Treasurer.

C. ELTON HINSHAW, *Chair*

Report of the Editor

American Economic Review

General Nature of the Editorial Process

The editorial process at the *Review* is a cooperative enterprise. Papers are received at the Princeton office and then distributed to the appropriate Co-Editor for a decision with respect to publication. Co-Editors arrange for refereeing and accept and reject papers in an entirely decentralized process. Only when a paper is accepted for publication is it sent again to the Princeton office for final editing and typesetting.

Historically the Editor and Co-Editors of the *Review* come from a breadth of fields designed to cover the largest substantive areas in economics from which we receive submissions. Co-Editors who specialize in the fields of macroeconomics and economic theory are essential to handle the many submissions in these areas. The remainder of our submissions come mainly from the various applied fields of microeconomics, and the two other Co-Editors have expertise in these areas.

The Co-Editing process has four important advantages. First, papers are generally assigned to a Co-Editor who has some substantive knowledge of the topic and research in the relevant field of economics. This is helpful both in the assignment of referees (as Table 6 indicates, the vast majority of submissions to the *Review* are submitted to peer refereeing) and in the decision whether to publish a submission.

Second, the co-editor process permits us to avoid the actual (or apparent) conflict of interest that results when an editor handles a colleague's paper. As a general rule, editors are never assigned papers written by authors at the same institution.

Third, the co-editor process provides a way by which we can handle the enormous number of manuscripts that we receive in a reasonably prompt fashion. It is inconceivable that a single editor could deal fairly and efficiently with the nearly 1,000 submissions that we receive annually.

Finally, the co-editing system permits a considerable amount of turnover in the editorial

TABLE 1—CO-EDITORS, *American Economic Review*
APRIL 1, 1985—PRESENT

Editor	Affiliation
1) Orley Ashenfelter	Princeton University
2) Robert H. Haveman	University of Wisconsin
3) John G. Riley	University of California—Los Angeles
4) John B. Taylor	Stanford University
5) Hal R. Varian	University of Michigan
6) Bennett T. McCallum	Carnegie Mellon University
7) Paul R. Milgrom	Stanford University
8) John Y. Campbell	Princeton University
9) Roger H. Gordon	University of Michigan
10) R. Preston McAfee	University of Texas
11) Kenneth D. West	University of Wisconsin
12) Dennis N. Epple	Carnegie Mellon University/ Northwestern University
13) Mark W. Watson (interim)	Princeton University
14) Matthew D. Shapiro	University of Michigan

process without major disruption in the handling of our submissions. This turnover tends to ensure that no single point of view dominates the criteria for acceptance of papers while it maintains an orderly and continuous editorial process.

Table 1 contains a list of the Co-Editors at the *Review* since the beginning of the co-editing process. About one Co-Editor is replaced at the *Review* each year.

The *Review*'s Board of Editors meets once a year for a discussion of broad policy matters, when that is necessary. (It was a vote of the Board, for example, that led to the *Review*'s decision to adopt double-blind refereeing.) The main purpose of the Board, however, is to assist in the refereeing of submissions, especially those papers (such as comments and replies) that require both tact and considerable effort in their handling. A list of the current

TABLE 2—BOARD OF EDITORS, *American Economic Review* APRIL 1, 1985—PRESENT

George A. Akerlof	Gene M. Grossman	John Roberts
Joseph G. Altonji	Daniel S. Hamermesh	Richard Roll
James E. Anderson	Gordon Hanson	David H. Romer
Alan J. Auerbach	Robert H. Haveman	Paul M. Romer
David K. Backus	Robert J. Hodrick	Thomas Romer
Kyle W. Bagwell	Kevin D. Hoover	Richard E. Romano
David P. Baron	R. Mark Isaac	Nancy L. Rose
Theodore C. Bergstrom	Adam B. Jaffe	Alvin E. Roth
Timothy J. Besley	George E. Johnson	David E. M. Sappington
Rebecca M. Blank	Paul L. Joskow	Richard L. Schmalensee
Robin W. Boadway	Kenneth L. Judd	Myron S. Scholes
Timothy F. Bresnahan	John H. Kagel	Suzanne A. Scotchmer
Charles C. Brown	John F. Kennan	Andrew R. Schotter
Clive D. Bull	Mervyn A. King	Matthew D. Shapiro
John Y. Campbell	Meir G. Kohn	Steven Shavell
H. Lorne Carmichael	Paul R. Krugman	John B. Shoven
Stephen G. Cecchetti	Karen K. Lewis	Kenneth J. Singleton
Lawrence J. Christiano	R. Preston McAfee	Robert S. Smith
Michael R. Darby	Bennett T. McCallum	Gary Solon
Steven N. Durlauf	Deirdre McCloskey	Barbara J. Spencer
George W. Evans	John McMillan	Jeremy C. Stein
Henry S. Farber	Paul R. Milgrom	Curtis R. Taylor
Marjorie A. Flavin	Robert A. Moffitt	Robert H. Topel
Robert P. Flood	Dale T. Mortensen	Richard Tresch
Jacob A. Frenkel	Maurice Obstfeld	Hal R. Varian
Timothy S. Fuerst	Edgar O. Olsen	W. Kip Viscusi
Don Fullerton	Christina H. Paxson	Carl E. Walsh
Jordi Galí	Wolfgang Pesendorfer	David N. Weil
Nancy Gallini	Robert H. Porter	Kenneth D. West
Claudia Goldin	Valerie A. Ramey	David W. Wilcox
Roger H. Gordon	Sergio T. Rebelo	John D. Wilson
Philip E. Graves	Jennifer F. Reinganum	Michael D. Woodford
Jo Anna Gray	Peter C. Reiss	Susan Woodward
Reuben Gronau	John G. Riley	Leslie Young

and past members (since 1985) of the Board of Editors is contained in Table 2.

Editorial Process

The editorial process has continued to work very smoothly during the past year. However, we are still catching up in the production process, which had been seriously delayed in 1996–1997 by the unanticipated and unplanned departure of our Managing Editor.

As Table 3 indicates, the number of submissions has remained at an all-time high level. This high level of submissions, coupled with our attempt to publish longer substantive articles, reduced the chance that a submitted paper will eventually be published to an all-time low level in 1997. Production delays were responsible for a further decline in the chances

of publication, but this should be offset in the next two years as we catch up in our production process.

Table 4 indicates that we published fewer articles and shorter papers in the *Review* in 1997 than in 1996, and that the total number of pages published decreased also. I and my Co-Editors have adopted a conscious policy of attempting to increase the number of major substantive articles we publish at the expense of shorter papers, comments, and replies. This policy had earlier been reflected in our publication statistics. The more recent decline in articles and pages published will be offset by an increase in pages published in the next two years.

Tables 5 and 6, when compared with the results for last year, indicate there has been little change in the speed with which we han-

TABLE 3—MANUSCRIPTS SUBMITTED AND PUBLISHED, 1978–1997

Year	Submitted	Published	Ratio, published-to- submitted
1978	649	108	0.17
1979	719	119	0.17
1980	641	127	0.20
1981	784	115	0.15
1982	820	120	0.15
1983	932	129	0.14
1984	921	138	0.15
1985	952	128	0.13
1986	987	123	0.125
1987	843	99	0.12
1988	844	100	0.12
1989	946	116	0.12
1990	911	100	0.115
1991	884	110	0.12
1992	950	108	0.11
1993	900	94	0.10
1994	953	91	0.10
1995	929	88	0.095
1996	976	85	0.087
1997	976	66	0.068

Note: The submissions reported for every year refer to the last two months of the previous year and the first ten months of the year reported.

dle manuscripts that are ultimately rejected (median time to rejection remained constant). A major cause of the delay in handling manuscripts can be attributed to the use of referees and to the slowness with which we receive reports. Table 6 shows this correlation between speed of rejection and the number of referees used.

Table 7 indicates that there has been some increase since last year in the speed with which accepted papers were published in the *Review*. Unlike rejected papers, accepted papers take longer in the revision process. The lag between acceptance and publication is primarily a result of technology and, in the last year, to unanticipated personnel changes. This lag should, therefore, decline in the next two years.

The subject matter distribution of papers published in the *Review* in 1996 and 1997 is contained in Table 8. It remains my impression that the distribution of published papers reflects fairly accurately the distribution of pa-

TABLE 4—SUMMARY OF CONTENTS, 1996 AND 1997

	1996		1997	
	Number	Pages	Number	Pages
Articles	52	1,020	44	848
Shorter Papers, including				
Comments and				
Replies	29	260	19	156
Announcements and				
Reports		38		42
Index and				
frontmatter		22		22
Total		1,340		1,068

pers submitted and has not changed much recently.

Papers and Proceedings

The 19th volume of the *Papers and Proceedings* to be prepared by the editorial staff of the *Review* appeared in May 1997. The past year this task has been very capably handled by Ronald Oaxaca (of the University of Arizona) and David Baldwin, our former Managing Editor. I am deeply indebted to both of them for the difficult work under extraordinarily tight deadlines that they have so capably performed.

Co-Editors and Board of Editors

I now edit the *Review* with the assistance of Dennis Epple (Carnegie Mellon University), Preston McAfee (University of Texas), and Matthew Shapiro (University of Michigan). I am deeply indebted to them for the

TABLE 5—DISPOSITION OF MANUSCRIPTS, 1996 AND 1997

	July 1, 1995– June 30, 1996	July 1, 1996– June 30, 1997
Manuscripts received	936	976
Completed processing	640	572
Accepted	10	17
Rejected	630	555
Currently in process	296	404

TABLE 6—DISTRIBUTION OF EDITORIAL DECISION LAGS BETWEEN RECEIPT AND REJECTION,
JULY 1, 1996–JUNE 30, 1997

Weeks to rejection	Total number of manuscripts	Percentage	No outside referees	One referee	Two referees	Three or more referees
0–4	19	3	11	5	3	0
5–6	33	4	0	12	20	1
7–8	55	7	1	10	38	6
9–10	70	9	0	18	40	12
11–12	73	9	0	8	47	18
13–14	79	10	0	9	41	29
15–16	69	9	0	6	44	19
17–21	129	17	0	10	70	49
22–26	86	11	0	5	54	27
27–30	54	7	0	1	31	22
31–35	43	6	0	0	22	21
36–52+	63	8	0	2	12	49
	773	100	12	86	422	253

conscientious effort they have expended over the last year. Also important in the publication of the *Review* is the work of Lynn Fleisher, Managing Editor, to whom I am also indebted. The Board of Editors now consists of 40 members, and I am indebted to them all for their efforts. Board members are selected to reflect the highest level of scholarship in the economics profession from the breadth of different fields represented in our submissions. More than fine scholarship is expected of a Board member, however. Board members are also selected because of their conscientiousness, good judgment, and professional reliability. When possible, we

like to select Board members from those economists who have been especially helpful in the outside refereeing process.

TABLE 8—SUBJECT MATTER DISTRIBUTION
OF PUBLISHED MANUSCRIPTS, 1996 AND 1997

Subject category	Published	
	1996	1997
General economics and teaching	0	0
Methodology and history of economic thought	0	0
Mathematical and quantitative methods	9	10
Microeconomics	22	9
Macroeconomics and monetary economics	14	13
International economics	7	10
Financial economics	3	6
Public economics	6	4
Health, education, and welfare	1	3
Labor and demographic economics	5	4
Law and economics	1	1
Industrial organization	4	2
Business administration and business economics; marketing; accounting	0	0
Economic history	1	0
Economic development, technological change, and growth	5	0
Economic systems	0	0
Agricultural and natural-resource economics	2	0
Urban, rural, and regional economics	1	1
Other special topics	0	0
	81	63

TABLE 7—AVERAGE PUBLICATION LAGS
BY JOURNAL ISSUE

Journal issue	Number of weeks lag		
	Receipt to acceptance	Acceptance to publication	Receipt to publication
March 1997	76	28	104
June 1997	109	36	145
September 1997	88	42	130
December 1997	110	38	148

TABLE 9—COPIES PRINTED, SIZE, AND COST OF PRINTING AND MAILING, 1997 AER

Issue	Copies printed	Pages		Cost		
		Net	Gross	Issue	Reprints	Total
March	28,911	258	280	\$ 69,518.32	\$1,332.03	\$ 70,850
May	28,822	522	544	115,048.20	4,900.58	119,949
June	28,751	224	248	61,745.04	1,283.20	63,028
September	28,581	330	360	81,017.29	1,715.09	82,732
December ^a	28,766	256	288	64,814.40	1,475.25	66,290
Annual miscellaneous ^b						15,000
Total:		1,590	1,720			\$417,849

^a Estimated.^b Estimated: based on costs of preparing mailing list, extra shipping charges, and storage costs of back issues.

Three members of the Board completed their terms during 1997: Alan Auerbach, Kyle Bagwell, and Lorne Carmichael. Four resigned their positions in view of pressing responsibilities elsewhere: David Backus, Rebecca Blank, Nancy Rose, and David Wilcox. I am most grateful to them and to the continuing members: James Anderson, David Baron, Theodore Bergstrom, Timothy Besley, Charles Brown, Stephen Cecchetti, Timothy Fuerst, Don Fullerton, Jordi Galí, Nancy Gallini, Gene Grossman, Gordon Hanson, Mark Isaac, Adam Jaffe, Paul Joskow, Karen Lewis, Deirdre McCloskey, Paul Milgrom, Robert Moffitt, Christina Paxson, Wolfgang Pesendorfer, Valerie Ramey, Sergio Rebelo, Jennifer Reinganum, Peter Reiss, Richard Romano, David Romer, David Sappington, Andrew Schotter, Gary Solon, Curtis Taylor, Kip Viscusi, Carl Walsh, David Weil, Kenneth West, and Michael Woodford. I am also grate-

ful for the assistance of many other associates who make it possible to edit and produce the *Review*. I am indebted to our office manager, Kathy Simkanich, who replaced Shirley Griesbaum when she retired after many years of dedicated service; and to our new editorial assistant, Irene Rowe, and Rita Boccanfuso for the fine work they have performed over the past year. I also thank the Co-Editors' secretaries: Carolyn Bartle (Matthew Shapiro's office); Patricia Niber (Dennis Eppler's office); and Judith Searcy (Preston McAfee's office).

As always, the published version of this report contains the list of referees who have volunteered their services during 1997. We extend our deepest appreciation for the time and energy they have devoted to the advancement of our science.

ORLEY ASHENFELTER, *Editor*

REFEREES

A. B. Abel
J. M. Abowd
D. Acemoglu
L. F. Ackert
A. Admati
P. Aghion
N. I. Al-Najjar
A. Alesina
C. R. Alexander
F. Allen
S. G. Allen
J. Alm

D. Altig
G. Anderson
J. E. Anderson
P. M. Anderson
S. P. Anderson
J. Andreoni
J. D. Angrist
M. Armstrong
R. J. Arnott
G. B. Asheim
S. Athey
A. Atkeson

S. E. Atkinson
O. P. Attanasio
A. J. Auerbach
D. Austen-Smith
L. M. Ausubel
C. N. Avery
R. B. Avery
I. Ayres
L. C. Babcock
P. Bacchetta
K. Back
D. K. Backus

M. Bagnoli
K. W. Bagwell
P. Bajari
G. Baker
J. B. Baker
L. C. Baker
M. Baker
G. S. Bakshi
R. E. Baldwin
B. Balk
D. Balkenborg
L. Ball

C. L. Ballard	G. J. Borjas	J. P. Caulkins	P. Debaere
D. P. Baron	D. Bos	S. G. Cecchetti	C. A. de Bartolome
G. F. Barrett	J. Boudoukh	M. Chang	G. L. Debelle
J. M. Barron	J. Bound	V. V. Chari	D. V. DeJong
E. J. Bartelsman	A. L. Bovenberg	J. Chavas	E. Dekel
K. Basu	M. Bowes	K. Y. Chay	U. Demiroglu
S. Basu	S. Bowles	Y. Che	R. A. de Mooij
M. Baxter	J. W. Boyd	L. Chen	V. Denicolo
M. R. Baye	M. Boyer	Z. Chen	M. Denny
C. M. Beach	J. C. Brada	J. A. Chevalier	M. Dewatripont
R. Beason	R. M. Braid	P. Chiappori	P. A. Diamond
A. Beltratti	J. A. Brander	G. Chichilnisky	F. X. Diebold
A. Ben-Ner	L. Branstetter	R. Chisik	J. DiNardo
R. Bénabou	R. A. Braun	B. R. Chiswick	G. Dionne
V. R. Bencivenga	Y. M. Braunstein	I. Cho	A. K. Dixit
H. D. Benjamin	R. A. Brecher	C. C. Chu	S. Djajic
P. Berck	T. F. Bresnahan	T. Chung	N. A. Doherty
A. N. Berger	J. A. Brickley	T. E. Clark	P. Dolton
J. Berger	S. G. Bronars	S. T. Coate	R. G. Donaldson
J. Bergin	C. C. Brown	J. H. Cochrane	S. Donnerfeld
P. R. Bergin	G. Brown	R. M. Coen	J. A. Doucet
T. C. Bergstrom	E. K. Browning	M. A. Cohen	B. E. Dowd
J. B. Berk	J. K. Brueckner	H. L. Cole	S. J. Dowrick
J. Berkowitz	M. Bryan	M. G. Coles	A. Drazen
B. S. Bernanke	J. Bryant	J. Conlisk	J. C. Driscoll
A. B. Bernard	S. Bucovetsky	P. J. Cook	M. Dudey
D. Bernhardt	J. Bughin	R. W. Cooper	S. N. Durlauf
B. D. Bernheim	L. T. Bui	T. E. Cooper	K. E. Dynan
S. T. Berry	K. Burdett	B. R. Copeland	D. Easley
G. Bertola	R. Burguet	K. Corts	J. Eaton
T. J. Besley	R. V. Burkhauser	P. J. Coughlin	N. Economides
H. Bester	A. C. Burnside	D. Cox	A. S. Edlin
G. Biglaiser	L. Busch	J. C. Cox	L. Eeckhoudt
S. Bikhchandani	K. F. Butcher	P. C. Cramton	I. Ehrlich
K. G. Binmore	D. A. Butz	J. Crémer	T. Eisenberg
G. Bittlingmayer	R. J. Caballero	M. Crew	T. Ellingsen
D. Black	B. Caillaud	K. J. Crocker	R. P. Ellis
R. Blair	S. M. Calabrese	R. Cronovich	E. Emch
R. M. Blank	C. F. Camerer	W. J. Crowder	W. E. Encinosa
F. D. Blau	M. R. Caputo	M. J. Crucini	C. Engel
F. Bloch	D. E. Card	J. Currie	M. Engers
G. C. Blomquist	J. A. Carlson	D. M. Cutler	W. B. English
R. W. Blundell	H. L. Carmichael	J. I. Daniel	D. Epstein
R. W. Boadway	W. J. Carrington	W. H. Dare	L. Epstein
P. Bohm	C. D. Carroll	A. F. Daughety	N. R. Ericsson
H. Bohn	R. T. Carson	C. Davidson	D. S. Evans
M. Boldrin	M. Carter	D. D. Davis	W. N. Evans
G. E. Bolton	A. C. Case	D. R. Davis	R. C. Fair
E. W. Bond	J. P. Caskey	S. J. Davis	R. E. Falvey
P. Boone	T. N. Cason	G. De Arcangelis	R. Färe
M. Bordinon	B. Cassiman	J. A. Dearden	R. E. Farmer
M. D. Bordo	J. Cassing	A. V. Deardorff	J. W. Faust

E. M. Feasel	R. Given	B. Hansen	C. Ichniowski
T. J. Feddersen	E. L. Glaeser	R. Hansen	S. Imrohoroglu
R. C. Feenstra	A. Glazer	G. H. Hanson	R. P. Inman
J. S. Feinstein	S. M. Gleason	M. S. Hanson	R. Innes
A. M. Feldman	G. Glomm	E. A. Hanushek	Y. M. Ioannides
R. Fernandez	J. Gokhale	J. D. Harford	P. N. Ireland
C. Fershtman	L. S. Goldberg	J. E. Harrington	R. M. Isaac
G. S. Fields	P. K. Goldberg	M. Harris	H. Itoh
D. N. Figlio	C. D. Goldin	R. M. Harstad	M. O. Jackson
M. G. Finn	C. Gollier	A. F. Haughwout	H. G. Jacoby
M. Fiorina	J. F. Gomes	D. M. Hausman	A. B. Jaffe
A. C. Fisher	A. C. Goodman	G. M. Heal	C. M. James
E. O. Fisher	T. J. Goodspeed	J. C. Heaton	E. Janeba
M. J. Fishman	E. Goodstein	J. F. Helliwell	P. Jehiel
C. J. Flinn	R. H. Gordon	M. Hellwig	C. Jencks
Z. Fluck	G. Gorton	E. Helpman	R. Jensen
C. L. Foote	L. H. Goulder	J. V. Henderson	U. J. Jermann
J. E. Foster	K. J. Graddy	K. Hendricks	A. John
R. Fowles	F. C. Graham	I. Henriques	G. E. Johnson
D. E. Frame	J. W. Graham	B. E. Hermalin	D. H. Joines
R. H. Frank	O. Grandville	J. Hersch	C. I. Jones
M. Freeman	J. S. Gray	G. D. Hess	C. M. Jones
S. Freeman	W. B. Gray	A. L. Hillman	G. Jonsson
D. Friedman	E. J. Green	J. R. Hines	P. L. Joskow
L. Froeb	R. C. Green	J. Hirshleifer	S. J. Kachelmeier
T. S. Fuerst	R. J. Green	R. A. Hirth	J. H. Kagel
D. Fullerton	S. M. Greenstein	J. P. Hoehn	C. M. Kahn
S. A. Gabriel	A. W. Gregory	W. L. Holahan	L. M. Kahn
F. Gahvari	T. A. Gresik	B. Holmstrom	G. L. Kaminsky
E. Gal-Or	D. M. Grether	C. A. Holt	R. Kanbur
I. L. Gale	Z. Griliches	D. M. Holthausen	E. Kandel
J. Galí	V. Grilli	D. Holtz-Eakin	E. J. Kane
N. Gallini	M. Grinblatt	H. J. Holzer	S. N. Kaplan
N. Gandal	J. T. Grogger	Y. Hong	L. Kaplow
T. Gao	T. Groseclose	K. D. Hoover	C. Karayalcin
M. R. Garfinkel	G. M. Grossman	H. A. Hopenhayn	E. Katok
N. Gaston	H. I. Grossman	H. Horn	H. Katz
S. Gates	M. Grossman	A. Hornstein	M. L. Katz
G. Gaudet	J. Gruber	I. J. Horstmann	C. Kazimi
M. S. Gaynor	T. W. Guinnane	M. T. Horvath	M. Keen
D. Genesove	J. Guo	P. Howitt	P. J. Kehoe
W. M. Gentry	J. Gyourko	E. P. Howrey	T. J. Kehoe
M. Gertler	D. D. Haddock	H. W. Hoynes	D. A. Kendrick
R. H. Gertner	K. M. Hagerty	D. Hsieh	E. L. Khalil
J. Geweke	P. A. Haile	G. Huberman	N. M. Kiefer
S. Ghatak	B. J. Hall	G. Hueckel	L. Kilian
M. S. Gibson	R. E. Hall	C. R. Hulten	M. R. Killingsworth
I. Gilboa	K. F. Hallock	D. Hummels	J. Kim
S. Gilchrist	Y. Hamao	L. Hunnicutt	M. S. Kimball
M. Gilligan	D. S. Hamermesh	W. Hurley	A. Kimhi
T. W. Gilligan	J. D. Hammond	T. A. Husted	S. Kimmel
V. A. Ginsburgh	T. H. Hannan	E. Huybens	T. C. Kinnaman

K. Kiyono	J. V. Leahy	A. Manning	S. Mongell
B. Klein	E. E. Leamer	M. Manove	J. D. Montgomery
M. W. Klein	B. Lee	D. Marcouiller	S. Moorthy
P. R. Kleindorfer	T. Lee	R. A. Margo	J. Morduch
P. J. Klenow	P. Legros	L. Martin	J. Morgan
S. Klepper	D. E. Leigh	K. E. Maskus	S. Morris
A. Klug	H. Leland	C. F. Mason	G. Moscarini
J. L. Knetsch	T. Lemieux	S. E. Masten	M. Motta
M. M. Knetter	D. Levin	S. Matthews	H. Moulin
T. J. Kniesner	R. Levine	S. J. Matusz	B. R. Moulton
B. H. Kobayashi	J. A. Levinsohn	P. Mauro	C. Mulligan
C. D. Kolstad	A. Levinson	W. Mayer	J. C. Murdoch
P. Kooreman	D. A. Levinthal	J. Mazumdar	R. J. Murnane
S. Kortum	S. D. Levitt	M. McAleer	D. Mustard
L. J. Kotlikoff	D. T. Levy	B. T. McCallum	G. M. Myers
D. Kovenock	P. I. Levy	P. S. McCarthy	A. Nayyar
S. Kozicki	A. Lewbel	D. McCloskey	D. A. Neal
L. Kranich	K. K. Lewis	K. E. McConnell	J. P. Neary
R. E. Kranton	T. R. Lewis	M. McCubbins	T. J. Nechyba
K. Krehbiel	E. Ley	J. McDermott	W. S. Neilson
M. Kremer	D. D. Li	R. McDonald	W. Nelson
K. Krishna	W. Li	D. McFadden	D. Neumark
K. F. Kroner	G. D. Libecap	K. McGarry	R. G. Newell
R. S. Kroszner	P. Lin	T. G. McGuire	E. Newlon
A. B. Krueger	B. L. Lipman	M. McKee	A. F. Newman
P. Krusell	J. M. Litwack	S. McLanahan	D. Newmark
J. V. Krutilla	A. Lizzeri	R. P. McLean	S. Ng
K. M. Krutilla	M. P. Loeb	J. McMillan	T. Nilssen
P. J. Kuhn	G. F. Loewenstein	S. G. Medema	E. Niou
P. Kumar	S. Lohmann	Y. P. Mehra	S. Nitzan
P. H. Kupiec	N. V. Long	J. Melvin	T. H. Noe
E. R. Kwerel	J. R. Lott	E. G. Mendoza	R. G. Noll
F. E. Kydland	D. J. Lucas	F. Menezes	W. D. Nordhaus
C. LaCasse	R. E. Lucas	A. Merlo	G. Norman
S. E. Lach	R. D. Ludema	M. Merz	S. C. Norrbin
J. M. Lacker	F. T. Lui	D. Messick	R. L. Oaxaca
J. Laffont	A. Lusardi	G. E. Metcalf	J. N. Ochs
J. T. LaFrance	R. Lutter	A. Metrick	P. G. O'Connell
E. L. Lai	A. B. Lyon	C. Meyer	S. A. O'Connell
D. Laibson	T. P. Lyon	T. J. Miceli	G. S. Oettinger
J. Laitner	M. Machina	R. Michaely	L. E. Ohanian
R. J. LaLonde	J. K. MacKie-Mason	I. Mihov	E. O. Olsen
P. Lam	W. B. MacLeod	G. M. Milesi-Ferretti	M. Olson
L. Lambertini	B. C. Madrian	P. R. Milgrom	J. A. Ordoover
V. E. Lambson	W. A. Magat	M. H. Miller	M. B. Ormiston
S. Landefeld	S. P. Magee	S. M. Miller	A. Orphanides
E. M. Landes	G. Maggi	C. Minter Hoxby	J. M. Orszag
M. Landsberger	V. Maksimovic	J. A. Miron	M. J. Osborne
K. Lang	J. M. Malcomson	K. Miyagiwa	A. O'Sullivan
P. Lasserre	B. G. Malkiel	P. Mizen	A. J. Oswald
J. P. Lawarrée	D. A. Malueg	R. A. Moffitt	M. Ottaviani
E. P. Lazear	A. Manelli	B. Moldovanu	A. L. Owen

P. Oyer	S. Ramaswamy	S. Rose-Ackerman	R. Sethi
S. E. Page	G. Ramey	H. S. Rosen	G. Shaffer
M. Paglin	V. A. Ramey	B. P. Rosendorff	C. Shannon
D. Pal	A. Rampini	H. Rosenthal	P. Shapiro
R. B. Palmquist	A. J. Randall	R. W. Rosenthal	W. W. Sharkey
J. C. Panzar	M. R. Ransom	S. S. Rosenthal	S. Shavell
L. E. Papke	R. H. Rasche	M. R. Rosenzweig	S. M. Sheffrin
S. L. Parente	E. B. Rasmusen	D. R. Ross	L. M. Sheiner
J. A. Parker	A. Ratfai	J. J. Rotemberg	P. A. Shively
I. W. Parry	D. Rathbun	A. E. Roth	J. F. Shogren
D. O. Parsons	D. J. Ravenscraft	R. Rothschild	L. D. Shore-Sheppard
B. P. Pashigian	S. T. Rebelo	C. E. Rouse	D. Showalter
M. V. Pauly	J. B. Rebitzer	B. R. Routledge	H. Sieg
C. H. Paxson	A. Redish	A. Rubinstein	J. G. Silber
N. Pearson	P. Regibeau	G. D. Rudebusch	R. D. Simpson
J. Peck	S. Reichelstein	A. Rustichini	J. L. Sindelar
P. Pecorino	C. Reimers	E. Sadka	N. Singh
J. Peek	J. F. Reinganum	M. Sadler	A. Siow
D. Peled	M. Reinsdorf	E. Sadoulet	S. Skaperdas
G. G. Pennacchi	P. C. Reiss	B. Salanie	C. Skiadas
J. Penrod	S. Reiter	D. Salehi-Isfahani	J. Skinner
E. C. Perotti	D. Reitman	M. K. Salemi	M. E. Slade
R. Perotti	J. D. Reitzes	J. Salerno	M. J. Slaughter
M. Perozek	P. Reny	G. Saloner	D. T. Slesnick
I. Perrigne	P. Rey	L. W. Samuelson	A. D. Slivinski
P. Perron	E. M. Rice	G. Sanchez	F. A. Sloan
M. Perry	J. Richard	A. Sandroni	M. Smart
M. K. Perry	M. D. Richardson	D. E. Sappington	T. M. Smeeding
W. Pesendorfer	W. Rieber	T. J. Sargent	A. A. Smith
M. Peters	R. G. Riezman	R. Sarin	B. D. Smith
G. Phillips	J. Ríos-Rull	T. R. Sass	J. P. Smith
G. Piga	J. A. Ritter	M. Sattinger	L. Smith
J. Piggott	R. Rob	S. Scandizzo	R. S. Smith
T. Piketty	J. Roberts	S. Schaefer	V. L. Smith
R. S. Pindyck	J. M. Roberts	H. Schaller	A. Snow
J. Pischke	M. J. Roberts	K. A. Scharf	C. M. Snyder
P. Pita Barros	R. Robertson	T. C. Schelling	J. M. Snyder
R. Pitchford	P. K. Robins	F. M. Scherer	J. Sobel
M. M. Pitt	A. J. Robson	H. Schlesinger	P. Soderlind
S. Polasky	J. Rochet	R. L. Schmalensee	B. L. Sohngen
R. V. Polavarapu	W. Roeger	K. M. Schmidt	S. J. Solnick
R. H. Porter	J. E. Roemer	R. Schob	G. Solon
J. M. Poterba	J. H. Rogers	J. K. Scholz	J. L. Solow
S. M. Potter	R. Rogerson	A. R. Schotter	R. M. Solow
C. Prendergast	W. P. Rogerson	J. L. Schrag	J. Sonstelie
D. Primont	G. Roland	M. Schwartz	B. Sopher
J. E. Prisbrey	R. Roll	M. Sefton	N. S. Souleles
Y. Qian	R. E. Romano	U. Segal	N. Spatafora
J. M. Quigley	C. D. Romer	K. Segerson	B. J. Spencer
M. Rabin	D. H. Romer	T. M. Selden	M. M. Spiegel
R. Radner	T. Romer	G. A. Selgin	Y. Spiegel
R. G. Rajan	N. L. Rose	R. Selten	D. F. Spulber

S. Srivastava	A. Tornell	M. Waldman	M. A. Williams
E. Stacchetti	J. S. Tracy	J. M. Walker	S. R. Williams
D. O. Stahl	D. Trefler	M. Walker	C. A. Wilson
F. Stahler	S. J. Trejo	N. E. Wallace	J. D. Wilson
R. W. Staiger	J. E. Triplett	C. J. Waller	H. Winter
O. Stark	P. K. Trivedi	C. E. Walsh	F. Wirl
R. N. Stavins	K. R. Troske	C. Wang	D. Wittman
C. Stefanadis	J. T. Tschirhart	H. Wang	B. L. Wolfe
M. Stegeman	D. Tsiddon	P. Wang	E. N. Wolff
J. C. Stein	G. Tullock	R. Wang	P. Wolfson
M. Stinchcombe	G. K. Turnbull	Y. Wang	A. Wolinsky
K. Storesletten	S. J. Turnovsky	M. Waterson	K. I. Wolpin
T. Stratmann	C. Udry	J. C. Watson	K. Wong
W. Strayer	H. Uhlig	L. Waverman	S. A. Woodbury
C. Stuart	T. S. Ulen	W. E. Weber	M. D. Woodford
G. Stuart	A. Ulph	K. Weigelt	T. Worrall
F. A. Sturzenegger	M. Ureta	D. N. Weil	R. Wright
M. J. Stutzer	S. Valdes-Prieto	D. E. Weinstein	Y. Wu
A. Sutherland	J. B. Van Huyck	E. R. Weintraub	X. Xing
K. Suzumura	T. Van Zandt	M. S. Weisbach	A. Yelowitz
N. R. Swanson	H. R. Varian	D. L. Weisman	J. M. Yinger
J. Swierzbinski	E. H. Veendorp	M. L. Weitzman	S. Yitzhaki
G. Tabellini	F. R. Velde	L. Welling	H. Yoshikawa
R. F. Tamura	A. J. Venables	A. J. Wellington	A. Zamouline
D. Tarr	S. F. Venti	S. H. Wellisz	R. J. Zeckhauser
A. M. Taylor	J. Ventura	Q. Wen	J. Zeira
C. R. Taylor	J. Vercammen	Y. Wen	J. F. Zender
L. J. Taylor	N. Vettas	I. Werner	T. Zha
M. S. Taylor	J. Vickers	K. D. West	H. Zhang
D. J. Teece	D. R. Vincent	J. Weymark	J. Zhou
L. L. Tesar	J. R. Vincent	K. Whelan	Z. Zhu
A. V. Thakor	G. Violante	M. J. White	J. P. Ziliak
R. H. Thaler	W. K. Viscusi	M. W. White	L. Zingales
J.-F. Thisse	X. Vives	T. M. Whited	P. M. Zorn
D. Thomas	P. P. Wakker	N. T. Wilcox	M. A. Zupan
T. H. Tietenberg	J. Waldfogel	D. E. Wildasin	M. Zurlinden
G. M. Tootell	D. M. Waldman	M. O. Wilhelm	

Report of the Editor

Journal of Economic Literature

The *Journal of Economic Literature's* mission is to help keep members of the Association informed of research developments in various fields of economics. This is accomplished by providing them with articles that describe and evaluate research progress on particular issues and by supplying a bibliographic guide to books, journals, and dissertations.

The *Journal's* work is shared between two offices. The Pittsburgh office is supervised by Drucilla Ekwurzel, and it has responsibility for the bibliographic departments including the contents of current periodicals, abstracts of papers, and book annotations. Mary Kay Akerman serves as an Assistant Editor, and Asatoshi Maeshiro as an Editorial Consultant. They are helped by Patricia Andrews, Elizabeth Braunstein, Ruby Glasgow, Amy Lawrence, Ann Norman, Douglas Quint, and Elizabeth Thornton, in addition to a staff of part-time workers. I am grateful to them for their valuable contributions throughout the year.

The Pittsburgh branch of the *Journal* is responsible for the annual *Index of Economic Articles* and the Economic Literature Index. In 1997, we brought out the 1993 *Index*, and in 1998 we hope to publish the 1994 and 1995 *Indexes*. The Economic Literature Index is available on line through Dialog Information Services and on compact disc through an arrangement with SilverPlatter Information. The EconLit database provides over 25 years of journal citations, book annotations, and dissertation titles. Since 1994, it includes the *Journal's* book reviews and Abstracts of Working Papers in Economics, the latter through an agreement with Cambridge University Press. A truncated version of EconLit (named EconLit-AEA) is available to members of the American Economic Association at an annual rate of \$75 (plus shipping and handling). It provides bibliographic information covering approximately the past 15 years. For 1998, we project sales of EconLit-AEA to about 800 members.

Each year, the Board of Editors reviews the set of journals whose contents are listed in the published version of the *Journal*. It also considers requests by new journals for listing. Last year, we dropped 59 journals and added five other journals to the set of those whose contents are listed in the printed version of the *Journal*. These decisions are consonant with the policy stated in previous Annual Reports of the Editor to move bibliographic information to the electronic database and to make the printed *Journal* more selective. Our goal is to contain the growth in costs and in the size of the *Journal*. Information on the articles of all journals we receive appears on the electronic versions of our databases.

Table 1 describes the movements in the *Journal's* pages since 1980. It indicates that, notwithstanding the persistent growth in the number and size of journals, we have been able to keep the pages devoted to Current Periodicals to a level about the same as that in 1983. The popularity of our electronic data bases makes it likely that the Board of Editors will continue to recommend that certain bibliographic information be available in electronic form only.

The Articles and Communications and the Book Review departments of the *Journal* are managed from Stanford University. During 1997, the *Journal* published 21 major articles and 172 book reviews. Alex Field supervises the Book Review Department, and Frank Wolak shares with me the editorial duties associated with reviewing the manuscripts. We receive excellent support from Anita Makler and Britt Ellis. I am most grateful to them for their very significant contribution to the production of the *Journal*.

In 1995, we introduced a CD-ROM version of the quarterly issues of the *Journal of Economic Literature*, the first journal in economics available in this form. The December CD-ROM contains all four issues of the *Journal* for the year. Members may choose to receive their issues of the *Journal* either in the conventional printed form or in the form of a CD-ROM that

TABLE 1—JEL PAGES BY DEPARTMENT, 1980–1997

Year	Articles and communications	Book reviews	New book annotations	Current periodicals	General index	Total
1980	366	294	276	1,072	26	2,034
1981	342	286	270	1,059	23	1,980
1982	331	251	300	1,069	23	1,974
1983	305	239	281	1,086	38	1,949
1984	354	225	314	1,193	37	2,123
1985	364	237	299	1,306	38	2,244
1986	326	250	308	1,343	41	2,268
1987	345	251	315	1,352	40	2,303
1988	419	241	318	1,240	40	2,258
1989	334	251	328	1,254	41	2,208
1990	323	234	366	1,339	43	2,305
1991	462	224	362	1,091	22	2,161
1992	754	226	412	1,169	24	2,585
1993	748	230	406	1,093	25	2,502
1994	533	276	446	1,117	28	2,400
1995	547	291	517	1,024	27	2,406
1996	507	264	484	1,195	27	2,477
1997	717	249	444	1,096	24	2,530

Notes: In 1987, the *Journal of Economic Literature* took over from the *American Economic Review* the responsibility of publishing the list of Doctoral Dissertations in Economics. This item is added to "Current Periodicals" which also includes the Contents of Current Periodicals, the Subject Index of Articles in Current Periodicals, and Selected Abstracts.

may be read by PC-DOS, Windows, Macintosh, or UNIX machines. Order forms for the CD-ROM version are available in every issue of the *Journal*. At the time of writing this report, about 3,500 members are choosing to receive their issues of the *Journal of Economic Literature* in the form of a CD-ROM.

I refer members to the statement of editorial objectives and policies set forth in an Editor's Note at the beginning of the March 1986 issue. In accordance with these policies, our articles are commissioned by the Editor. The Editor welcomes proposals for and outlines of such articles. An outline should be about four pages in length plus another page or two (but not more) consisting of those references that are likely to occupy an important place in the proposed article. We look for articles that explain and evaluate the issues in a major research endeavor. We do not seek encyclopedic surveys of the literature, especially those that have the appearance of a long sequence of abstracts strung together.

In the assessment of manuscripts and proposals of papers, we have benefited greatly from the services of referees, some of whom prepared outstanding reports. I thank them

very much indeed for their advice. The quality of our articles is largely the product of our very conscientious referees. The referees we used in 1997 are listed at the end of this report.

The Editor also commissions book reviews. Our policies with respect to book reviews are contained in a statement in the March 1992 issue of the *Journal*. We now reproduce this statement at the beginning of our Book Review Department in each issue of the *Journal*.

I thank all members of the Board of Editors for their work on the *Journal's* behalf during the year. Many have made very considerable contributions to the *Journal's* product. Alan Auerbach, Peter Howitt, Peter Reiss, F. M. Scherer, and John Whitaker have kindly agreed to serve another term on the Board. A. B. Atkinson, Anne Case, John Roemer, and Alan Stockman are leaving the Board. I thank them very much indeed for their extremely helpful advice and assistance over the years. I shall be proposing to the AEA Executive Committee that Lewis Evans, James Levinsohn, Glenn Loury, Jennifer Reinganum, Michael Rothschild, Suzanne Scotchmer, Hans-Werner Sinn, Hal Varian, and I join the Board of Editors next year.

In January 1998, I step down as Editor of the *Journal*. In 12 years of editing this journal, I have accumulated heavy debts. Very many economists have served as outstanding referees and have written most thoughtful and professional reports on manuscripts I have sent them. Authors have patiently responded to my objections to their papers and have rewritten their articles to deal with the Editor's idiosyncratic concerns. I am sure there are authors who have gnashed their teeth after reading my letters. The Nashville office of the American Economic Association has been very supportive of my efforts. The people there work with uncommon courtesy and quiet, rare, efficiency. Drucilla Ekwurzel and her staff at the Pittsburgh office of the *Journal* have been a delight to work with. Their efforts on behalf of the Association are really remarkable. Alex

Field's shepherding of the Book Review section of the *Journal* has been outstanding. Frank Wolak has shared the load of managing the manuscripts, and he has done so with cheerfulness and proficiency. The people who have really run the Stanford office of the *Journal* are Britt Ellis and Anita Makler. I thank them most sincerely for their devoted and very effective efforts. They have modestly allowed people to believe that I, not they, have been Editor of the *Journal*.

The new Editor of the *Journal* is John McMillan. It passes to him with my best wishes for a successful editorship. I am confident the *Journal* will thrive under his stewardship.

JOHN PENCAVEL, *Editor*

REFEREES

H. J. Albers	L. Evans	J. Lerner	P. A. Samuelson
J. Alm	M. Fafchamps	R. Levine	W. Schulze
D. Audretsch	D. Foley	A. Lewbel	T. Sicular
M. Baxter	R. H. Frank	B. J. Loasby	M. E. Slade
T. Bayoumi	J. Gans	G. F. Loewenstein	J. B. Slemrod
A. H. Beller	M. Gertler	J. G. MacKinnon	D. Southgate
A. Ben-Ner	D. Ghura	M. McClellan	R. W. Staiger
P. Berck	H. Gintis	E. S. Mills	P. E. Stephan
J. Bhagwati	P. K. Goldberg	R. J. Murnane	T. Stoker
H. Binswanger	C. Goldin	B. J. Naughton	J. M. Swinkels
R. M. Bird	P. Gottschalk	T. J. Nechyba	D. J. Teece
M. L. Blackburn, Jr.	V. Hajivassiliou	D. Netzer	J. Temple
L. E. Blume	J. C. Haltiwanger	J. P. Newhouse	L. L. Tesar
A. Booth	P. Hammond	J. Niehans	R. H. Thaler
J. A. Brander	G. K. Helleiner	M. Olson, Jr.	M. R. Tool
P. L. Brock	J. V. Henderson	A. Ortmann	A. Velasco
D. J. Brown	J. Hirshleifer	M. M. Pitt	A. D. Velenchik
E. Burmeister	R. Jackman	I. Png	M. Wallerstein
R. D. Cairns	J. P. Jacobsen	R. Pollak	E. G. West
C. Camerer	A. B. Jaffe	R. E. Quandt	L. E. Westphal
M. Clerici-Arias	J. Janssen	M. Rabin	D. E. Wildasin
J. Conlisk	C. I. Jones	G. Ramey	B. Wilkinson
I. A. Coxhead	P. Joskow	J. Reinganum	D. Winch
C. Cummins	A. Kochar	J. G. Riley	B. L. Wolfe
S. Dowrick	J. A. Krautkraemer	P. K. Robins	M. Woodford
J. Driffill	T. Kuran	S. Rose-Ackerman	G. Wright
T. Dunne	N. R. Lardy	T. J. Rothenberg	
G. D. Ellison	E. Leeper	A. B. Royalty	
P. D. Evans	T. Lemieux	M. H. Rutherford	

Report of the Editors

Journal of Economic Perspectives

The tenth full year of publication for the *Journal of Economic Perspectives* has been a productive one. The *Journal* has continued its pattern of publishing a mixture of symposia, individual papers, features, correspondence, and other material. In 1997, the *Journal* included eight symposia: the natural rate of unemployment, wage inequality, the distribution of world income, European unemployment, electronic journals in economics, the European single currency, fiscal federalism, and telecommunications deregulation around the world. These were complemented by articles on a wide range of topics, including: capacity utilization and the self-serving bias; electricity deregulation and Paul Samuelson's *Economics* text; and inflation targeting and Austrian economics. In addition, the *Journal* continued publication of several features. Bernard Saffran's "Recommendations for Further Reading" continued to appear in each issue. Joseph Persky managed the "Retrospectives" feature on topics in the history of economic thought. Eugene Steuerle of the Urban Institute handled the "Policy Watch" column. Greg Duncan of Northwestern University oversaw the "Data Watch" columns. Charles Holt of the University of Virginia administered the "Classroom Games" feature. Richard Thaler of the University of Chicago continues his "Anomalies" column. As in years past, the *Journal* continued to publish "Correspondence" and "Notes." The editors are taking a more active role in encouraging correspondence to the *Journal* and intend to increase the number of letters and responses that will be published.

The 1997 issues of the *Journal* included 896 pages, consisting of 41 regular articles and 13 feature articles, plus Correspondence, Notes, and miscellaneous items like advertisements and announcements. The total number of pages was 2-percent lower than the annual average over the previous four years. Table 1 provides a breakdown of how the pages were allocated, together with comparisons for the years since 1993.

At the end of 1997, the Associate Editors who have completed their terms are Henry Aaron,

Brookings Institution; Francine D. Blau, Cornell University; Anne C. Case, Princeton University; Gregory Mankiw, Harvard University; Frederic S. Mishkin, Columbia University; and Suzanne Scotchmer, University of California–Berkeley. We thank each of them for their efforts and for their contributions to the *Journal* over the past three years. The incumbent Associate Editors are: David Colander, Middlebury College; Robert Gibbons, Cornell University; Oliver Hart, Harvard University; Peter Murrell, University of Maryland; Joseph Newhouse, Harvard University; Dani Rodrik, Harvard University; Bernard Saffran, Swarthmore College; Richard Schmalensee, Massachusetts Institute of Technology; and Gavin Wright, Stanford University. The expiration dates for the current terms of the ongoing Associate Editors are listed in Table 2.

Although most articles appearing in the *Journal* are normally solicited by the Editors and Associate Editors, a number of unsolicited proposals and papers arrive directly at the *Journal*'s administrative office. The *Journal* charges no submission fee and thus attracts a wide range of proposals. Most submissions are inappropriate for the *Journal* for one reason or another, often including level of specialization, style of exposition, or narrowness of focus. Others are good ideas that overlap to some extent with plans that have already been made. Still other suggestions offer possibilities and spark a discussion between the Editors, the Associate Editors, and the author which sometimes leads to a *JEP* article. The *Journal* typically receives between 150 and 200 unsolicited proposals each year. Given the number of solicited articles and the *Journal*'s space and budgetary limitations, no more than a handful of the unsolicited proposals typically end up appearing in the *Journal* in a given year. While the *Journal* remains open to ideas and input from all sources, it remains the case that the overwhelming proportion of articles appearing in the *Journal* are solicited by the Editors and Associate Editors.

One recent innovation at the *Journal* is to hold some of our symposia live each year, thanks to a grant from the Andrew W. Mellon Foundation. The Symposium on the Natural

TABLE 1—PAGE DISTRIBUTION FOR THE *Journal of Economic Perspectives*, 1993–1997

	Number of pages				
	1993	1994	1995	1996	1997
Total	920	896	960	880	896
Full-length articles (number of articles)	694 (40)	684 (36)	722 (39)	670 (37)	700 (41)
Introductions and comments (number of articles)	49 (7)	50 (3)	54 (5)	12 (2)	—
Features (number of features)	96 (11)	64 (8)	113 (12)	120 (13)	104 (11)
Correspondence	19	36	20	20	26
Notes	24	24	23	22	26
Table of contents	8	8	8	8	8
Advertisements and announcements	30	30	20	28	32

Rate of Unemployment, which appeared in the Winter 1997 issue, was held at the Georgetown University Conference Center in Washington, DC, in September 1996, while the Symposium on Discrimination in Product, Credit, and Labor Markets, which will appear in the Spring 1998 issue, was held live at Princeton University in June 1997. The *Journal* benefits from such live symposia in several ways: greater visibility among economists who can attend the live symposia; a broader array of discussion, feedback, and criticism for the authors; and a heightened incentive for first drafts to arrive on time.

The editorial team of the *Journal* remained stable in 1997. Alan B. Krueger of Princeton University and J. Bradford De Long of the University of California–Berkeley continued as Editor and Co-Editor of the *Journal*, respectively. The administrative operations of the *Journal* continue to be located at the Hubert H. Humphrey Institute of Public Affairs at the University of Minnesota. We thank the Humphrey Institute for administrative,

computer, and logistical support, and for providing a good home for the *Journal*.

The job of Editorial Associate was held by Carmen Largaespada until July 1997. For the three years that she worked at *JEP*, Carmen's combination of intelligence, persistence, attention to detail, and good humor made a substantial contribution to the smooth functioning of the *Journal*. Since July, the job of Editorial Associate has been held by Melinda Prescher, who is making a strong start at filling the sizable shoes of her predecessor.

Timothy Taylor continued in 1997 as Managing Editor of the *Journal*. As in years past, the Editors feel that they cannot overstate the role that he has played in the operation of the *Journal*. He has managed the day-to-day operations of the *Journal* smoothly and ensured that the fundamental objectives of the *Journal* are satisfied. He has performed the difficult task of persuading authors to amend and rewrite their articles with vigor, verve, and skill, and he has shown that it is possible to edit papers in such a way as to increase their clarity and accessibility, while still retaining the distinctive voice of each author.

Requested action: Approval of Associate Editors starting in 1998, to serve a three-year term expiring at the end of 2000, to be announced at the Executive Committee meeting in January 1998.

TABLE 2—SCHEDULED EXPIRATION OF CURRENT TERMS FOR ASSOCIATE EDITORS

End of 1998	End of 1999
Gibbons	Colander
Murrell	Hart
Newhouse	Rodrik
Saffran	
Schmalensee	
Wright	

ALAN KRUEGER, *Editor*
J. BRADFORD DE LONG, *Co-Editor*

Report of the Director

Job Openings for Economists

The total number of new jobs listed increased from 1,613 last year to 1,879 this year. Academic jobs increased by 115, from 1,039 to 1,154; nonacademic ones increased by 151, from 574 to 725. Table 1 shows total listings (employers), total jobs, new listings, and new jobs by type for each issue of *JOE* in 1997.

Table 2 shows the number of employers by category (four-year colleges, universities with graduate programs, federal government, etc.) for each of the seven issues. Academic institutions continue to be the major advertisers—about 65 percent of the total number of employers listing vacancies.

Table 3 shows the number of listings by field of specialization. The *JEL* category (C), Mathematical and Quantitative Methods, led in popularity, dropping Money and Macro (E) to the second spot. International (F) came in third, edging out Microeconomics (D), which came in fourth. Industrial Organization (L) finished fifth. These five categories have consistently dominated the lists since *JOE* was first published.

JOE is available on-line through the Internet. There is no charge. Anyone can search its data base by keyword or the classification system of the *Journal of Economic Literature*. The on-line *JOE* is available soon after the publication of each new issue. Only the most recent version is maintained on-line. Since going on the Internet three years ago, *JOE* has lost almost 65 percent of its subscribers. In December 1994, there were

TABLE 1—JOB LISTINGS FOR 1997

Issue	Total listings	Total jobs	New listings	New jobs
<i>Academic:</i>				
February	59	115	53	93
April	33	75	28	46
June	31	67	31	67
August	44	88	41	83
October	197	414	185	386
November	135	271	135	271
December	147	322	91	208
Subtotal	646	1,352	564	1,154
<i>Nonacademic:</i>				
February	33	76	26	44
April	28	73	24	53
June	40	101	34	79
August	35	84	28	61
October	78	200	73	181
November	59	184	59	184
December	71	215	39	123
Subtotal	344	933	283	725
Total	990	2,285	847	1,879

about 3,100. Now there are 1,114. I expect the decline to slow but continue.

Violet Sikes does everything necessary to get *JOE* printed and mailed on time, including dealing with hostile callers who have missed the publication deadline. Employers and applicants are in her debt, as am I.

C. ELTON HINSHAW, *Director*

TABLE 2—NUMBER AND TYPES OF EMPLOYERS LISTING POSITIONS IN JOE DURING 1997

Issue	Four-year colleges	Universities with graduate programs	Federal government	State/local government	Banking or finance	Business or industry	Consulting or research	Other	Total
February	22	37	7	3	2	3	16	2	92
April	13	20	6	—	3	2	16	1	61
June	8	23	7	2	5	8	16	2	71
August	8	36	5	1	9	1	17	2	79
October	61	136	12	3	17	10	31	5	275
November	40	95	18	2	9	7	21	2	194
December	44	103	13	1	13	9	31	4	218
Totals	196	450	68	12	58	40	148	18	990

TABLE 3—FIELDS OF SPECIALIZATION CITED: 1997

Code	Fields	February	April	June	August	October	November	December	Totals
A	General Economics and Teaching	7	7	7	3	19	12	14	69
B	Methodology and History of Economic Thought	1	0	0	0	3	3	4	11
C	Mathematical and Quantitative Methods	28	22	21	26	104	67	76	344
D	Microeconomics	19	12	18	29	91	47	63	279
E	Macroeconomics and Monetary Economics	25	17	19	23	99	49	58	290
F	International Economics	19	21	9	21	85	55	72	282
G	Financial Economics	15	16	15	19	59	52	65	241
H	Public Economics	13	4	10	17	50	33	41	168
I	Health, Education, and Welfare	7	16	8	10	26	25	34	126
J	Labor and Demographic Economics	10	7	9	13	49	37	40	165
K	Law and Economics	7	2	4	2	11	7	11	44
L	Industrial Organization	16	10	20	24	83	47	72	272
M	Business Administration; Business Economics; Marketing, Accounting	6	6	8	7	20	11	18	76
N	Economic History	2	2	1	1	5	7	5	23
O	Economic Development Technological Change	14	7	7	10	38	29	35	140
P	Economic Systems	2	1	3	2	10	1	7	26
Q	Agricultural and Natural Resource Economics	24	16	18	14	39	31	33	175
R	Urban, Rural, and Regional Economics	7	4	5	7	25	20	16	84
Z	Other Special Topics	2	0	2	4	4	6	5	23
AF	Any Field	14	6	9	11	65	41	44	190
ZZ	Administrative positions	10	2	1	3	7	9	8	40
	Totals	248	178	194	246	892	589	721	3,068

Note: Fields of specialization codes are from the *Journal of Economic Literature*.

Report of the Committee on Economic Education

During May 2–4, 1997, the Committee sponsored its second active-learning workshop at the University of North Carolina at Chapel Hill. The workshop included sessions on the learning-theory case for active learning, discussion-leading, the case method, cooperative learning, the use of nontraditional writing assignments, and assessment of active-learning outcomes. The workshop staff included Patrick Conway (University of North Carolina), Michael Salemi (University of North Carolina), Phillip Saunders (Indiana University), Ann Velenchik (Wellesley College) and William Walstad (University of Nebraska). The workshop was attended by 39 participants who rated the workshop highly, judging it to be a better use of their time than their next best alternative.

The Committee sponsored a one-day workshop on January 4, 1998 as part of the Association program for the Allied Social Science Association Meetings. For the second year in a row, the one-day workshop was well attended and rated highly by participants. The Committee will continue to offer teaching workshops at the ASSA meetings as long as interest remains high.

Funds for the 1996–1997 Active Learning Project were generously provided by the Calvin K. Kazanjian Economics Foundation. The Workshop Program Directors, Michael Salemi and William Walstad, are currently seeking funds to offer new residential workshops and to offer one-day workshops at the annual meetings of several regional economics associations.

The Committee sponsored two sessions at the January 1998 ASSA meetings. The first comprised papers on the teaching of statistics and econometrics to undergraduate students. Interest in these papers was very high, and session attendance exceeded 150. The papers appear elsewhere in this volume. Our second sponsored session comprised papers on the use of experiments in the teaching of undergraduates. It, too, was well attended, and those interested in the papers may contact Denise Hazlett at the Department of Economics, Whitman College, Walla Walla, WA 99362, or by e-mail at hazlett@whitman.edu.

The Committee continues to track the recent cycle in the number of economics bachelor degrees. Now that another year's worth of data have become available, it has become clearer that the trough in degrees was reached in the 1995–1996 academic year.

Margo and Siegfried (*Journal of Economic Education*, Fall 1996, p. 328) found that the number of economics degrees experienced a local peak in 1990–1991. In that year, economics degrees accounted for 3.6 percent of the bachelor-degree total, a much higher share than the 1948–1994 sample average of 2.2 percent. For the next five years, the number of economics degrees fell precipitously. Economics degrees in 1995–1996 equaled only 73 percent of the 1990–1991 total.

Based on 111 responses to the AEA Universal Academic Questionnaire for 1997, John Siegfried estimates that the number of degrees awarded in economics was about 3.7-percent higher in 1996–1997 than in 1995–1996. My own calculations using enrollment and degree data for the University of North Carolina at Chapel Hill predicts a substantial additional increase in the number of degrees conferred during the 1997–1998 academic year. Siegfried's most recent findings are scheduled to appear in the Summer 1998 issue of the *Journal of Economic Education*.

On the occasion of his retirement from Indiana University, the Committee would like to commend Phillip Saunders for his many years of service to economic education. No one has had a more important impact on economic education in the past 25 years than Phil. We join many, many others in thanking him for his contributions and leadership.

Information about the Committee on Economic Education and its activities is now available on the World Wide Web by connecting to the American Economic Association web page at www.vanderbilt.edu/AEA and choosing the "Committee" option. To contact the Committee on Economic Education, please email Michael Salemi at Michael_Salemi@unc.edu.

MICHAEL K. SALEMI, *Chair*

Report of the Committee on the Status of Minority Groups in the Economics Profession

The primary objective of the Committee on the Status of Minority Groups in the Economics Profession (CSMGEP) is to increase the representation of minority groups in economics. Targeted minority groups are blacks, Hispanics, and Native Americans. In 1997, the major AEA-sponsored activity undertaken to facilitate achievement of this objective was the Summer Training Program, and the associated AEA Minority Scholarship.¹ CSMGEP also undertook a major fund-raising effort in 1997. The result has been to secure a significant four-year grant from the John D. and Catherine T. MacArthur Foundation to support the "Economics Pipeline Project," which will include the Summer Program as well as a series of new initiatives. This report provides some information about minorities in economics, an update on the Summer Program, and a description of the proposed Pipeline Project.

Minorities in Economics: An Overview

There continue to be relatively few blacks, Hispanics, and Native Americans in the economics profession. The small number of minority economists is closely related to the fact that few minority students pursue and are awarded doctoral degrees in economics. The National Opinion Research Center maintains data on the numbers of doctoral degrees conferred on U.S. citizens, by field and ethnic group. Relevant figures for 1976–1996 are shown in Table 1. On average, only 11 new economics Ph.D.'s per year were awarded to minorities during 1976–1978. This figure rose to an average of 18 per year during 1979–1981. The three-year average fluctuated between 16 and 20.7 during 1979–1993. Over the most recent three years, the average was 22.3. The timing and persistence of the post-1979 increase in minority economics doctoral

degrees is consistent with the hypothesis that the AEA Summer Program (discussed further below) had a positive and lasting effect.

As a percentage of total economics doctorates awarded to U.S. citizens, those to minorities rose from 1.9 percent during 1976–1978 to 5.5 percent during 1994–1996. However, much of the increase during 1976–1993 is attributable to the well-known decline in the denominator, reflecting the fall in nonminorities pursuing advanced economics degrees. Table 1 also shows trends in numbers of Ph.D.'s awarded by ethnic group.

AEA Summer Program for Minority Students

Since 1974, the AEA has sponsored an annual Summer Program for Minority Students which seeks to increase the number of individuals who pursue economics Ph.D.'s. Each summer, the Program provides roughly 20 of the strongest minority undergraduates who express interest in careers in economics with an intensive eight-week course of instruction in analytic materials essential to graduate study. Many of the minority economists now enrolled in graduate programs or already active in the profession benefited from their participation.

The Program is designed to be run by a host institution for a period of 3–5 years, after which time a new host is selected. The University of Texas at Austin (UT) has just completed its second year as Summer Program host. The 1996 and 1997 Programs, directed by Don Fullerton, appear to have been very successful. We note that the restructuring of the program in 1997 to separate the Training component from the Minority Scholarship component appears to have successfully addressed all relevant legal issues. UT's term as host has recently been extended from three years to five years—through 2000.

We are pleased to report that CSMGEP has secured adequate funding for the Summer Program for the period 1998–2001. Indeed, fund-raising was the Committee's major activity over the past two years. First, at the end of

¹ Unfortunately, the AEA Federal Reserve System (FRS) Dissertation Fellowship Program has been suspended.

TABLE 1—REPRESENTATION OF MINORITIES AMONG CONFERRED ECONOMICS DOCTORATES

Year	Economics doctorates					
	Total	Minority	Percentage minority	Black	Hispanic	Native American
1976–1978	571.3	11.0	1.9	7.7	2.7	0.7
1979–1981	519.3	18.3	3.5	7.3	9.0	2.0
1982–1984	464.0	19.3	4.2	11.0	8.0	0.3
1985–1987	442.3	16.0	3.6	6.3	8.3	0.4
1988–1990	419.0	20.7	4.9	10.0	10.0	0.7
1991–1993	386.7	20.0	5.2	12.3	7.3	0.3
1994–1996	407.0	22.3	5.5	12.0	10.0	0.3

Notes: Minority is defined as black, Hispanic, or Native American. Figures are numbers of degrees conferred, unless indicated otherwise.

Sources: National Research Council, Summary Report (1994: *Doctorate Recipients from U.S. Universities*, p. 78 (1976–1978). National Opinion Research Center, *Affirmative Action* (table 2: Ph.D.'s awarded to U.S. citizens by race/ethnicity) (1979–1996).

1996 we received a three-year grant (1997–1999) from the National Science Foundation (NSF), jointly funded by Research Opportunities for Undergraduates and the Economics Program. Second, at the end of 1997, we received a four-year grant (1998–2001) from the MacArthur Foundation which includes funding for the Summer Program. On this basis, NSF has indicated its willingness to extend its support through 2001. The extension is to include an increase in the total NSF grant. Third, we have raised some support from corporate sources. Finally, the AEA has agreed to an increase in its annual contribution to the Summer Program.

AEA Summer Minority Program Follow-up

CSMGEP together with staff at the University of Texas continue their efforts to locate and survey past participants in the Summer Program. Improving our information about what has happened to these individuals will enable us to better evaluate this initiative, and to strengthen the program as well as our other initiatives. We stress that this effort is still in progress, so that the data collected so far are incomplete. In particular, our tracking effort started with recent graduates.

Table 2 provides summary results of the information we have received so far. As shown, a total of 615 students participated in the Program since it was started in 1974. Of these, we

have surveyed 234 individuals. Data for each of the seven host institutions (with the relevant years of operation) are provided on separate lines. Not surprisingly, our data are most complete for recent participants from the UT and Stanford Programs.

Table 2 shows that 93 (40 percent) of the past participants surveyed actually enrolled in economics Ph.D. programs. Of particular note, 43 (48 percent) of the surveyed 89 participants from the Program at Stanford enrolled in Ph.D. programs.² We have located 24 past participants who have received doctorates in economics. Fifty individuals are still enrolled in graduate programs. However, 19 individuals left their graduate programs either with a terminal master's degree, or with no degree. We hope to reduce this dropout rate through a series of new interventions discussed below. Our data also show that 25 (11 percent) of the 234 students entered economics master's programs, and 70 (30 percent) entered other graduate programs, including business and law.

² The UT data should be interpreted with caution. Many of these participants are still completing their undergraduate degrees. In addition, some UT participants had been accepted to graduate programs in the previous spring.

TABLE 2—AEA SUMMER MINORITY PROGRAM FOLLOW-UP INFORMATION

Host institutions	Total for hours	Total surveys completed	Began Ph.D. program	Received Ph.D.	Received M.A. (terminal)	Still enrolled in program	Left program without degree
Texas (1996–1997)	41	41	11	0	0	10	1
Stanford (1991–1995)	123	89	43	1	5	32	5
Temple (1986–1990)	137	35	16	4	3	8	1
Wisconsin (1983–1985)	87	20	4	4	0	0	0
Yale (1980–1982)	88	29	11	8	2	0	1
Northwestern (1975–1979)	117	15	7	6	1	0	0
UC–Berkeley (1974)	22	5	1	1	0	0	0
Totals:	615	234	93	24	11	50	8

Source: See text.

The Economics Pipeline Project

As stated above, CSMGEP recently secured funding for its Economics Pipeline Project from the MacArthur Foundation. The objective of the Pipeline is to expand the pool of minority Ph.D. economists using a series of interventions targeted at critical junctures in their training and professional development. The initiative will include three interrelated Programs: an Outreach Program, the (existing) Summer Training Program, and a Mentor Program.

The Outreach and Mentor programs are new components, designed to build on the existing Summer Program so as to establish a pipeline, or longer-term support system, for minority students interested in pursuing economics Ph.D.'s. The Outreach Program will extend support by attracting students early in their college experience. It will seek to expose them to the range of options available to professional economists, and to provide them with information about pursuing such careers. The Mentor Program will extend support beyond the Summer Program's preparation for graduate school. Participating students will be linked to a group of professional economists and expected to maintain active contact. Mentors will be expected to work cooperatively

with the student's departmental adviser. These students and their mentors will participate in annual two-day workshops with formal and informal sessions on research and on succeeding in graduate school.

Combined with the existing Summer Program, these two new Programs will focus on helping students to navigate critical stages in their professional development, including interest in economics, preparation for graduate school, successful completion of core theory and field exams, and initiation of dissertation research.

Committee Membership

The 1997 Committee on the Status of Minority Groups in the Economics Profession was composed of seven members: Susan M. Collins (Chair; Georgetown University and The Brookings Institution), Lynn C. Burbridge (Rutgers University), Alvin E. Headen, Jr. (North Carolina State University), Willene A. Johnson (Federal Reserve Bank of New York), Barbara J. Robles (University of Texas), Richard Santos (University of New Mexico), and Warren C. Whatley (University of Michigan).

SUSAN M. COLLINS, *Chair*

Report of the Committee on the Status of Women in the Economics Profession

The American Economics Association (AEA) has charged the Committee on the Status of Women in the Economics Profession (CSWEP) with monitoring the position of women in the profession and with undertaking activities to improve that position. This report presents information on the position of women graduate students and faculty in academic economics departments and reports on the committee's activities during 1997.

The Hiring and Promotion of Women Economists in Ph.D.-Granting Departments

For the past three years, CSWEP has worked on developing its contacts in all of the Ph.D.-granting departments in the United States. One of the tasks of the CSWEP representatives in these institutions is to report on the status of women in their departments. CSWEP has been able to acquire more complete and accurate data than are available currently through the AEA Universal Academic Questionnaire (UAQ) which is mailed to all department chairs each fall. CSWEP sent out a questionnaire in September 1996 and was able to obtain information from 98 of its 120 contacts in comparison to the UAQ which received responses from 74 Ph.D.-granting economics departments in 1996.¹

Information from the CSWEP Questionnaire on the Status of Women Faculty.—Table 1 provides information on the share of women faculty at various ranks in the 98 Ph.D.-granting departments. Column (i) provides information on all 98 departments, while Column (ii) and (iii) provide information from the top 10 and 20 schools.

Table 1 indicates that the share of women with academic appointments in 1996 at the

Ph.D.-granting institutions decreases with rank. The growing group of nontenured faculty in economics departments consists disproportionately of women. Compared to the 24 percent of women receiving Ph.D.'s, of those faculty in non-tenure-track positions, 50.2 percent are women. Untenured tenure-track assistant professors are 23.8-percent female. Untenured associate professors are 9.1-percent women. Tenured associate professors are 15.4-percent women, and tenured full professors are 8.4-percent female. Among the top 20 schools, the numbers are lower at every rank, indicating less representation of women on the faculty in the very top-ranked departments, except in the tenured associate professor ranks of which 16.1 percent are women. The top 10 departments have higher percentages of untenured assistant and tenured associate professors who are women. The percentage of tenured full professors is 5.3 percent.

Information from the CSWEP Questionnaire on the Status of Women Graduate Students in Economics.—The availability of women to the economics profession depends on the pipeline of women being trained in economics. Table 2 reports information on women in graduate programs in economics, taken from the CSWEP 1996 questionnaire. For the academic year 1996–1997 about 30.5 percent of the first-year class are female. Slightly over 28 percent of those who are “ABD” (all but dissertation) are female. Yet only 24.1 percent of those receiving a Ph.D. in economics are female at the 98 Ph.D.-granting departments reporting.² The represen-

¹ CSWEP's sample contains only U.S. economics departments, while that of the AEA UAQ includes a few non-U.S. economics departments.

² A consistent series on the share of women Ph.D.'s in economics is obtained from the National Science Foundation's Annual Survey of Earned Doctorates. The National Science Foundation reports that 22.4 percent of the doctorates granted in economics in 1996 went to women, slightly less than CSWEP identifies. Information on two of the top 20 schools, however, is missing from the CSWEP data.

TABLE 1—SHARE OF WOMEN (PERCENTAGE) BY RANK,
PH.D.-GRANTING DEPARTMENTS, FALL 1996

Rank	(i) All Ph.D.- granting	(ii) Top 10	(iii) Top 20
Non-tenure track	50.2	45.5	50.0
Assistant professor (Untenured)	23.8	21.1	18.2
Associate professor			
Untenured	9.1	0.0	0.0
Tenured	15.4	20.0	16.1
Full professor (Tenured)	8.4	5.3	5.5

Source: Data collected by CSWEP, 98 of 120 Ph.D.-granting schools reporting in column (i), 9 out of 10 reporting in column (ii), and 19 out of 20 reporting in column (iii).

tation of women at the top 20 departments is very similar to that for all graduate departments. Approximately 30 percent of the entering class are women, 26 percent of the ABD's are women, and 22.7 percent of the Ph.D.'s are women. The percentage for the top 10 graduate programs is slightly less favorable for women. While the percentage of new Ph.D.'s who are women has improved since the inception of CSWEP in 1972, the percentage of new Ph.D.'s in economics is relatively low when compared to the 22 fields reported by the National Science Foundation in 1995.

Information from the CSWEP Questionnaire on the Job Market Facing Women.—Table 3 shows how women fared in the job market in 1996 relative to men. With approximately 24 percent of the Ph.D.'s going to women, only 20 percent of the academic jobs at Ph.D.-granting departments went to women, and 26 percent of the jobs at non-Ph.D.-granting departments went to women. At the top 20 schools, women received 22.7 percent of the degrees and 19.2 percent of the jobs at Ph.D.-granting departments. These women received a disproportionate share of the jobs at non-Ph.D.-granting departments, 42.3 percent. These data suggest that women from the top schools are going to smaller private or state institutions rather than continuing their careers at Ph.D.-granting departments. Moreover, a disproportionate share of women did not find jobs in 1996.

The Committee's Activities

CSWEP Ongoing Activities.—CSWEP is involved in a wide range of activities to help bring women into the profession and to increase the rates at which women are promoted at various stages of their careers. As part of its ongoing efforts to increase the participation of women on the AEA program, CSWEP organized six sessions for the January 1998 ASSA meetings, three on gender-related topics and three on women, risk, and the financial markets. In addition, we organized a roundtable discussion, "Social Security Reform: How Will Women Fare?" to highlight the important effect that recent changes will have on the economic position of women. CSWEP also holds a business meeting at the annual meetings to report to associates about its activities and to hear from the AEA membership suggestions for future activities. To support junior women meeting senior women, a hospitality suite is staffed by members of the Committee.

New CSWEP Initiatives.—This year's meetings are particularly important for CSWEP. We celebrate the 25th anniversary of its founding. To honor the occasion several new initiatives came on line. First, a newly formatted newsletter was designed and produced, and it debuted with a special anniversary edition for Fall 1997. The newsletter contained articles on the progress of women in academe and business. A new mission

TABLE 2—SHARE OF WOMEN (PERCENTAGE) AMONG PH.D. STUDENTS AT DIFFERENT POINTS OF ACADEMIC PROGRESS, 1996–1997 SCHOOL YEAR

Academic progress	(i) All Ph.D.- granting	(ii) Top 10	(iii) Top 20
First year	30.5	26.5	30.2
ABD	28.3	23.9	26.4
Ph.D.	24.1	18.6	22.7

Source: Data collected by CSWEP, 98 of 120 Ph.D.-granting schools reporting in column (i), 9 out of 10 reporting in column (ii), and 19 out of 20 reporting in column (iii).

statement was passed by the Committee during its September meeting and was published, reiterating its commitment to the original goals of CSWEP. The newsletter also contained an article about the past, present, and future goals of CSWEP. CSWEP's website has been redesigned from the pilot effort of last year. Visitors to the new site will find navigating the options more user-friendly and the contents more informative.

At this year's meetings is the first NSF/AEA-CSWEP workshop to team-mentor junior women economists. CCOFFE: Creating Career Opportunities for Female Economists is a two-day workshop that brings together eight senior women economists and 40 junior women economists from the top universities in the country to work cooperatively on each other's projects as teams. In addition, there are sessions on publishing, grant-writing, networking, and balancing life choices. Similar workshops will be conducted at the regional meetings this year. By the end of the year NSF/AEA-CSWEP will have increased the chances of 200 women getting tenure within the next six years.

CSWEP's Regional Activities.—To assist women in the profession who cannot make it to national meetings, CSWEP organizes sessions at the Eastern, Southern, Midwest, and Western Economic Association meetings. As at the national meetings, sessions are on gender-related research and non-gender-related fields to showcase younger women economists. CSWEP is increasing its efforts to broaden the base of its organization by encouraging a closer liaison between the regional

governing boards and the formation of regional CSWEP committees to attend to the work of the region associations. In addition, CSWEP will conduct regional adaptations of the CCOFFE workshops at these meetings this year.

CSWEP's Network.—CSWEP has maintained its recently organized network of representatives at 120 Ph.D.-granting schools in the country. These representatives help the Committee monitor the progress of women at these schools and collect the information upon which elements of this report are based. This year we assisted the Committee on the Status of Minorities in the Economics Profession by expanding CSWEP's survey to include questions about race and ethnicity.

A Few Words of Thanks

The Committee thanks several people who have made major contributions to its effort. Joan Haworth, the Membership Secretary, and her staff maintain the Roster, send out annual membership reminders, and create customized listings for potential employers. In addition this year they have helped us redesign the website to bring their operation on line.

Two members left the Committee at the end of 1997: Maureen Cropper (The World Bank) and Kenneth Small (University of California-Irvine). Both of these members of the Committee did more than their fair share of the work over the past three years. They organized sessions at the national meetings, hosted the Committee in Washington, and co-edited the newsletter with me. Both members always did

TABLE 3—SHARE OF WOMEN (PERCENTAGE) PLACED IN JOB BY TYPE OF JOB,
AMONG STUDENTS ON THE JOB MARKET, WINTER AND SPRING 1996

Rank	(i) All Ph.D.-granting	(ii) Top 10	(iii) Top 20
U.S. Ph.D.-granting	20.2	19.6	19.2
U.S. other academic	26.4	30.8	42.3
U.S. public sector	20.5	21.1	32.5
U.S. private sector	28.0	25.0	25.9
Non-U.S. academic	21.1	12.0	9.8
Non-U.S. nonacademic	16.7	20.0	20.0
No job found	28.0	28.9	31.2

Source: Data collected by CSWEP, 98 of 120 Ph.D.-granting schools reporting in column (i), 9 out of 10 reporting in column (ii), and 19 out of 20 reporting in column (iii).

more than they needed to do and were always happy to do so.

Finally, CSWEP thanks Sally Scheiderer for keeping the Committee and all of its paper and cyber work on track. Denison University, and in particular the Department of Economics, the Department of Women's Studies, and the Laura C. Harris Chair, has contributed to the work of CSWEP with space, paper, telephones, and postage. Fi-

nally, CSWEP thanks Mary Winer and her staff at the AEA offices for their help and assistance. Marlene Height also has been a tremendous help with the logistics of setting up the CCOFFE workshop. All of these people have been wonderful to work with, and the Committee could not have done its work without their commitment.

ROBIN L. BARTLETT, *Chair*

Report of the Representative to the Social Science Research Council

Under a grant from the MacArthur Foundation, the SSRC is starting a new economics training initiative for Ph.D. students. The purpose of the program is to broaden training in economics by encouraging graduate students to undertake projects that involve new topics, are inter-disciplinary, or involve reexamination of standard assumptions. Students in economics Ph.D. programs will apply and be selected competitively to participate in the program starting after the first or second year of graduate school. The intent is that the program both ease students' transition from coursework to research and encourage students to select research topics for their dissertations which will broaden the frontiers of economics. The initial stage will involve participation in a summer program where leading economists from a variety of universities and specialties will give presentations concerning how they apply analytic tools to the investigation of critical social problems. Following the workshop, students in the program will have opportunities to apply for research assistantships, specialized training at other universities, or internships at institutions such as the Congressional Budget Office, the World Bank, or a data-collection organization such as the National Opinion Research Center. Further workshops will be arranged so that the stu-

dents in the program will become part of a continuing cohort and will benefit from advice from faculty at other universities in addition to their thesis advisers at their home universities.

The Steering Committee for the program includes Alan Blinder (Chair), George Akerlof, B. Douglas Bernheim, Dennis Carlton, Claudia Goldin, Paul Krugman, and William Nordhaus. The program will hold its first workshop during the summer of 1998. Further information concerning application procedures and deadlines can be obtained from David Weiman at the SSRC or from <http://www.ssrc.org>.

Another program of the SSRC which is interdisciplinary, but involves a strong economics component, is the Higher Education Research Program. It consists of a series of workshops on specific aspects of the higher-education industry, including recent workshops on research universities and on economies of scale/scope in higher education. The goal of the program is to encourage new research on the higher-education sector. David Weiman of SSRC is in charge of the program, and Roger Noll is a member of the Steering Committee.

MICHELLE J. WHITE, *Representative*

Report of the Representative to the National Bureau of Economic Research

The National Bureau of Economic Research studies a wide variety of economic issues. Much of this research is conducted by economists working individually, but a large part is organized into special projects. To disseminate the results of this research, during 1997 the NBER issued 442 working papers, published seven books and two NBER journals, prepared special issues of other journals, and circulated the monthly *Digest* and the quarterly *Reporter*. In addition, the NBER held numerous conferences and workshops, including the four-week-long Summer Institute.

Programs.—The NBER's ongoing programs generally meet twice during the academic year and once, for a longer period, during the Summer Institute.

Bureau programs (with directors in parentheses) are Aging (David Wise), Asset Pricing (John Campbell), Corporate Finance (Robert Vishny), Children (Janet Currie), Development of the American Economy (Claudia Goldin), Economic Fluctuations and Growth (Robert Hall), Health Care (Alan Garber), Health Economics (Michael Grossman), Industrial Organization (Nancy Rose), International Finance and Macroeconomics (Jeffrey Frankel), International Trade and Investment (Robert Feenstra), Labor Studies (Richard Freeman), Monetary Economics (N. Gregory Mankiw), Productivity (Zvi Griliches), and Public Economics (James Poterba).

Projects.—NBER's projects generally bring a dozen or more researchers together to work on a common topic. The projects' findings are usually distributed initially as NBER Working Papers, and final versions are often published as Bureau books. During 1997 the following projects were underway: The Economics of Education, Empirical Studies of the American Economy, Market Microstructure, Globalization and U.S. Labor Markets, International Capital Flows, and Privatizing Social Security.

Summer Institute.—During the summer of 1997, 920 economists attended the Summer

Institute. Approximately 300 papers were presented in 30 separate sessions.

Working Papers.—NBER Working Papers are circulated to libraries of economics departments and business schools around the world and are available for \$5 per title (plus a \$10 postage and handling charge per order for foreign requests). Academic subscriptions to the entire series are available for \$820 per year inside the United States. Information on foreign subscriptions or partial subscriptions to Working Papers in selected programs is available from the NBER Publications Department.

Books.—During 1997 seven new volumes were issued by the NBER: *The Microstructure of Foreign Exchange Markets* (Jeffrey Frankel, editor), *East Asian Seminar on Economics* (Takatoshi Ito and Anne Krueger, editors), *Reducing Inflation: Motivation and Strategy* (Christina Romer and David Romer, editors), *Health and Welfare During Industrialization* (Richard Steckel and Roderick Floud, editors), *In Pursuit of Leviathan* (Lance Davis, editor), *Differences and Changes in Wage Structure* (Richard Freeman, editor), and *The Effects of U.S. Trade Protection and Promotion Policies* (Robert Feenstra, editor).

Journals.—The NBER publishes two journals, which appear once each year. Ben Bernanke and Julio Rotemberg edited the *Macroeconomics Annual*, and James Poterba edited *Tax Policy and the Economy*, Volume 12. Subscriptions for these two journals can be ordered from the MIT Press.

In addition, the NBER sponsored a special issue of a journal. Jeffrey Frankel organized the International Seminar on Macroeconomics and edited the papers presented there for the *European Economic Review*.

Periodicals.—The monthly NBER *Digest* provides brief summaries of working papers of general interest. The *Reporter* contains longer summaries of recent research, abstracts of all working papers, and announcements of the publication of NBER books.

Both of these periodicals are available upon request.

Fellows.—During 1997, David Laibson of Harvard University visited the Bureau as an Aging Fellow.

Martin Feldstein continued as President of the NBER during 1997. Further information

on NBER activities is available in the *Reporter* or from Martin Feldstein, NBER, 1050 Massachusetts Avenue, Cambridge, MA 02138-5398.

JOHN J. SIEGFRIED, *Representative*